Universitätsklinikum Ulm, Klinik für Allgemein- und Viszeralchirurgie
Ärztliche Direktorin: Prof. Dr. D. Henne-Bruns
AG Versorgungsforschung
Leiter: Prof. Dr. Franz Porzsolt

Entwicklung eines Verfahrens zur Evaluierung von Screeningstudien anhand einer vielbeachteten Studie zum Lungenkarzinom-Screening aus dem Jahre 2011

Dissertation
zur Erlangung des Doktorgrades der Medizin
der Medizinischen Fakultät der Universität Ulm

vorgelegt von
Marie-Christin Rösch
Ulm

2015

Amtierender Dekan: Professor Dr. Thomas Wirth

1. Berichterstatter: Professor Dr. Franz Porzsolt

2. Berichterstatter: Professor Dr. Christel Weiß

Tag der Promotion: 12. Mai 2016

Für meine Eltern

Inhaltsverzeichnis

Abkürzungsverzeichnis

ACRIN:	American College of Radiology Imaging Network
AJR:	American Journal of Roentgenology
BMJ:	British Medical Journal
CASP:	Critical Appraisal Skills Programm
CCI:	Comission of Chronic Illness
CONSORT:	Consolidated Standards of Reporting Trials
CT:	Computertomographie
EVT:	Endpoint Verification Team
FDG:	Fluordesoxyglucose
FN:	False Negative
FP:	False Positive
HTA:	Health Technology Assessment
ICER:	Incremental Cost-Effectiveness Ratio
JAMA:	Journal of the American Medical Association
LDCT:	Low-Dose Computertomographie
LHR:	Likelihood Ratio
LHS:	Lung Health Study
LSS:	Lung Screening Study
NCI:	National Cancer Institute
NEJM:	New England Journal of Medicine
NETT:	National Emphysema Treatment Trial
NLHR:	Negative Likelihood Ratio
NLST:	National Lung Screening Trial
NNS:	Number needed to screen
NPW:	Negativer Prädiktiver Wert
NSCLC:	Non-Small Cell Lung Cancer
PAP-Abstrich:	Test zur Früherkennung von Gebärmutterhalskrebs nach Papanicolaou
PET:	Positronen-Emissions-Tomographie
PLCO:	The Prostate, Lung, Colorectal, and Ovarian Randomized Trial
PLHR:	Positive Likelihood Ratio
PPW:	Positiver Prädiktiver Wert

PRISMA:	Preferred Reporting Itmes for Systematic Reviews and Meta-Analyses
PSA:	Prostata-spezifisches Antigen
QALY:	Quality adjusted life year
QUADAS:	Quality Assessment of Diagnostic Accuracy Studies
RCT:	Randomized Controlled Trial
ROC-Kurve:	Receiver Operating Characteristic-Kurve
SCLC:	Small Cell Lung Cancer
STARD:	Standards for Reporting of Diagnostic Accuracy
Sv:	Sievert
TN:	True Negative
TP:	True Positive
USP:	Usability of scientific publication
WHO:	World Health Organization

1. Einleitung

1.1 Screeningprogramme in der heutigen Gesellschaft

„An ounce of prevention is worth a pound of cure.“ (Benjamin Franklin, 1736)

Screeningprogramme stellen in der heutigen westlichen Welt ein etabliertes Instrument dar, um Krankheiten möglichst früh zu diagnostizieren und um eine erfolgversprechende Therapie einzuleiten. Das Wort Screening ist sowohl in den Medien als auch in der Bevölkerung in aller Munde. Berühmte Persönlichkeiten werben in TV Serien und während Werbeunterbrechungen für verschiedene Vorsorgeuntersuchungen. Zum Beispiel setzte sich Mariele Millowitsch für das Mammographie-Screening ein und die Felix Burda Stiftung wirbt unter anderem mit Annette Frier 2011 und Günter Netzer 2009 für die Darmkrebsvorsorge.

Der Tod durch eine Krankheit wird als ein Geschehen angesehen, welches aus dem Blickwinkel der Bevölkerung durch die Änderung des Lebensstils und durch die Früherkennung oder Verhinderung von malignen Prozessen beeinflusst und verhindert werden kann. Die Medien wie auch unsere Gesundheitspolitik veranlassen die Menschen zu dem Glauben, dass durch einen gesunden Lebensstil und vor allem auch durch die regelmäßige Teilnahme an Vorsorgeuntersuchungen jeder Mensch die Chance besitzt, ein langes Leben zu führen. Diese gefühlte Sicherheit, dass jeder Mensch, der regelmäßig an Früherkennungsuntersuchungen teilnimmt vor gefährlichen Krankheiten bewahrt oder von diesen geheilt werden kann, spiegelt sich in der positiven Resonanz für Screeninguntersuchungen wider.

Das Zitat von Benjamin Franklin ist aktueller als jemals zuvor. Auch der Chirurg Charles Plumley Childe verkündete in seinem Buch „The Control of a Scourge, or How Cancer is Curable“ (1906), dass Krebs heilbar sein könnte, wenn er nur früh genug erkannt werden würde [50]. Nach seinem Sinne wird bis heute versucht, durch eine möglichst frühe Diagnose einer Erkrankung, eine Heilung von dieser zu

ermöglichen. Die moderne Medizin ist in der Lage, Risikogruppen in der Gesellschaft heraus zu kristallisieren, um so die Erfolgschancen auf richtig positive Ergebnisse zu erhöhen und dieses Klientel mit einer passenden und wirkungsvollen Therapie zu behandeln.

Bereits im Jahre 1922 wurde eine offizielle Empfehlung zur Gesundenuntersuchung von der amerikanischen Ärzteschaft unter dem Titel „the periodic health examination“ herausgegeben [15]. Es kann angenommen werden, dass zu Beginn die ausschlaggebende Motivation der Ärzte an diesen Untersuchungsreihen zu partizipieren nicht die innere Überzeugung der Effizienz dieser Maßnahme, sondern der finanzielle Anreiz war. Fast 30 Jahre später, 1951 wurde die CCI (Comission of Chronic Illness) Konferenz für präventive Maßnahmen bei chronischen Erkrankungen abgehalten. Dabei wurde Screening als mutmaßliche Erkennung von bislang unbekannten Krankheiten durch die Durchführung von Tests, Untersuchungen oder anderen Verfahren, die zügig angewendet werden konnten, definiert. Zudem wurde betont, dass ein Screeningtest nicht als ein diagnostisches Mittel angesehen würde [103].

Heutzutage ist es für den Laien und selbst für den praktizierenden Mediziner schwer zu differenzieren, aus welchen Beweggründen Präventions- und Screeningprogramme etabliert werden. Wer die tatsächlichen Nutznießer und wer die Verlierer solcher Programme sind, lässt sich kaum sagen. Der Patient, der im Mittelpunkt stehen und am meisten Nutzen aus diesen Untersuchungen ziehen sollte, gerät zusehends ins Abseits und andere Akteure wie die Krankenkassen, pharmazeutische oder medizintechnische Firmen erobern und fordern für sich die besten Plätze im Wettbewerb um die Gesundheit. In diesem undurchsichtigen System sind vor allem Ärzte auf aussagekräftige Studien angewiesen, die den Nutzen und die klinische Relevanz einer neuartigen Methode belegen und die eine verlässliche Quelle für sie darstellen sollte.

1.2 Präventionsprogramme und politische Entscheidungen

Auch auf politischer Ebene wurde 2005 ein Vorschlag für einen Gesetzesentwurf

zur Regelung von Präventionsprogrammen vorgelegt. Dieses Präventionsgesetz sollte die Bereiche Früherkennung (sekundäre Prävention) und Verhinderung (primäre Prävention) von Krankheiten sowie die Gesundheitsförderung abdecken. Es wird versucht, der Prävention auch auf politischer Ebene eine größere Bedeutung zu widmen und im Rahmen dessen auch das finanzielle Budget in diesem Bereich zu erhöhen. Der Gesetzgeber erwartet von den Krankenkassen mehr Angebote für Präventionsprogramme und die Einbindung der Kommunen zur Förderung von Maßnahmen im Sinne der Prävention. Dieser und weitere Gesetzesentwürfe wurden bisher allerdings abgelehnt, zuletzt im Herbst 2013. Als Gründe gelten, die fehlende Verknüpfung mit einer Antikorruptionsregelung und die ausschließliche Beteiligung der gesetzlichen Krankenkassen an den Gesetzesentwürfen [18][102].

1.3 Screening

Wenn man im Concise Oxford Dictionary unter „to screen“ nachschlägt, findet man unter anderem Folgendes: etwas aussieben, absieben, durchsieben oder auch etwas filtern. Dementsprechend wird mithilfe des Screenings versucht, Personen mit einem bestimmten Merkmal (Parameter, der für die Erkrankung spricht) zu finden beziehungsweise herauszufiltern. Zusätzlich wird eine Zielpopulation benötigt, die ein erhöhtes Risiko trägt an einer bestimmten Erkrankung zu leiden. Dieser Punkt gestaltet sich jedoch oftmals schwierig, da zum einen die Kausalität zwischen Risikofaktor und Erkrankung bekannt sein muss und zum anderen es meist mehrere Risikofaktoren für eine Erkrankung gibt, die zum Teil bis zum Zeitpunkt des Screenings noch nicht vollständig bekannt sind.

Die Prävention teilt sich in drei Kategorien auf, die voneinander klar abzugrenzen sind, als da wären die primäre, die sekundäre und die tertiäre Prävention. Die Primärprävention setzt sich zum Ziel durch eine gezielte Änderung von Verhaltensweisen und Gewohnheiten von Menschen, diese vor einer Erkrankung zu bewahren (zum Beispiel soll die Aufgabe des Rauchens die Erkrankung an

Lungenkrebs verringern). Bei der Sekundärprävention wird hingegen versucht, eine bereits bestehende Erkrankung möglichst früh zu diagnostizieren, um eine Heilungschance zu erhöhen. Maßnahmen, die versuchen das Fortschreiten oder die drohenden Komplikationen einer Krankheit zu verhindern, werden unter der tertiären Prävention zusammengefasst [98]. Auf die Bedeutung und Unterschiede dieser drei Kategorien werde ich in dem Kapitel Material und Methoden ausführlicher eingehen. Als gemeinsame Merkmale weisen Screening-Untersuchungen Folgendes auf: Die untersuchten Personen weisen bis dato keinerlei Merkmale auf, die auf die untersuchte Krankheit schließen lassen, das heißt sie sind in Bezug auf diese Erkrankung bisher gesund. Als weiterer Punkt steht das übergeordnete Ziel, den betroffenen/erkrankten Personen einer Intervention zuzuführen oder sie in Bezug auf ihre Erkrankung, falls keine Therapie möglich ist, im zukünftigen Lebensabschnitt vorzuwarnen (zum Beispiel beim Morbus Huntington) [70].

Für die Bewertung von neuen Screeningmaßnahmen werden randomisiert kontrollierte Studien benötigt, die eine Überlegenheit der neuartigen Screeningmethode gegenüber den bisher verwendeten Verfahren zur Diagnosestellung der gesuchten Erkrankung zeigt. Die Studienergebnisse werden anhand von bestimmten Kriterien bewertet, die zur Validität zusammengefasst werden. „Unter Validität einer wissenschaftlichen Aussage ist zu verstehen, dass diese auch tatsächlich bestätigt, was zu bestätigen sie vorgibt, (...)“ [64, S. 36]. Die Validität wird in eine interne und externe Validität aufgeteilt. Die interne Validität wird beeinflusst, wenn Bias (systematische Fehler) in einer Studie das Studienergebnis in eine bestimmte Richtung verändern. Bei der externen Validität ist die Reliabilität des Tests und der damit verbundenen Ergebnisse gefährdet, weswegen erwartete positive Effekte eines Screenings außerhalb der Studie nicht reproduziert werden können.

Bevor Maßnahmen für die Implementierung eines Screenings getroffen werden, sollten die in Frage kommenden Screeningverfahren ausführlich überprüft werden, sodass nicht Jahre später, wie beispielsweise beim Mammografie-Screening der Nutzen eines bereits etablierten Screeningtests angezweifelt wird. Die Zeitschrift

“Der Spiegel” brachte im Juli 2014 einen fünf-seitigen Artikel mit dem Titel “Unsinn in bester Qualität” heraus, indem kritisch über den Nutzen des Brustkrebs-Screenings berichtet wurde. In diesem Artikel wurde aufgezeigt, dass mehrere Studien herausgefunden hatten, dass zwar weniger Frauen an Brustkrebs starben, wenn sie zum Screening gingen, jedoch lebten diese durchschnittlich nicht länger, da sie an einer anderen Erkrankung (zum Beispiel Erkrankung des Herz-Kreislaufsystems) sterben würden. Es wurde betont, dass das Brustkrebs-Screening das erste Screening in Deutschland war, „(…) das auf ausdrücklichen Wunsch der Politik eingeführt wurde (...)” [26, S.101] und nicht durch Mediziner oder anderen Mitgliedern des Gesundheitswesens. Des Weiteren war die Gesamtsterblichkeit bei vergleichbaren Gruppen mit und ohne Screening gleich hoch, lediglich in der Screeninggruppe starb von 1000 Frauen eine innerhalb von zehn Jahren weniger an Brustkrebs. Dies lässt sich prozentual als eine relative Reduktion der Mortalität an Brustkrebs von 20% beschreiben. Zudem wird von den Cochrane-Wissenschaftlern angenommen, dass fünf von 1000 gescreenten Frauen an Brustkrebs behandelt wurden, obwohl dieser Tumor ihr Leben zu keiner Zeit beeinträchtigt hätte. Dieses Screening kostete die gesetzlichen Krankenkassen im Jahre 2012 220 Millionen Euro. Dies führt zu einem Ressourcenverbrauch an Gesundheitsleistungen, die für andere Untersuchungen und Therapien fehlen. Ein ähnliches Bild findet sich beim PSA-Screening, weshalb diese Maßnahme zur sekundären Prävention umstritten ist und derzeit nicht mehr empfohlen wird [26].

1.4 Ziel der Arbeit

In meinen Recherchen habe ich festgestellt, dass es hervorragende Empfehlungen und Paper zur Validitätsanalyse von klinischen Studien gibt, jedoch sind vergleichbare Instrumente für Screeningstudien wesentlich schwieriger zu finden. Darüber hinaus werden die Anleitungen zur Validitätsprüfung von Screeningstudien mit denen von Diagnostikstudien zusammengefasst, obwohl sich beide Verfahren erheblich unterscheiden.

Unsere Arbeitsgemeinschaft beschäftigt sich mit verschiedenen Screeningmethoden, wie zum Beispiel dem Glaukom-Screening, PSA-Screening und dem Lungenkarzinom-Screening. Bei unserer Bearbeitung der unterschiedlichsten Screening- und Diagnostikstudien ist uns aufgefallen, dass erhebliche Unterschiede zwischen den einzelnen Screeningverfahren bestehen und dass zum anderen bedeutende Differenzen zwischen Screening- und Diagnostikstudien existieren.

Ein weiteres Thema, dass ich in meiner Arbeit abhandeln möchte, besteht in der Nachforschung, inwiefern Medizinstudenten in Bezug auf Screeningverfahren und die Einschätzung der Wertigkeit dieser ausgebildet werden. In diesem Zusammenhang stellten die Auswertung des medizinischen Curriculums und die Überprüfung der Inhalte medizinischer Standardlehrwerke der Inneren Medizin wichtige Kerngebiete meines Vorgehens dar.

Ziel meiner Arbeit ist es, anhand eines konkreten Beispiels und zwar dem Screening auf Lungenkarzinom, die Kriterien zu benennen, die erforderlich sind, um die Validität einer Screeningstudie zu bewerten. Diese Kriterien werden anschließend in einem zweiten Schritt zu einer Fragensammlung zusammengefasst, um Lesern einer Screeningstudie ein Instrument anzubieten, mit dessen Hilfe sie die Validität und klinische Relevanz einer Studie und des geprüften Tests evaluieren können. Dieses Verfahren wird auf die Studie aus dem New England Journal of Medicine (NEJM) aus dem Jahre 2011 mit dem Titel “Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening” angewendet, wobei die Studie und der Screeningtest überprüft werden und ausführlich auf die Grenzen und positiven Aspekte der neuen Methode hingewiesen wird.

2. Material und Methoden

Zunächst soll verifiziert werden, ob das Thema Screeningmethoden und die Bewertung dieser im Ausbildungscurriculum der Medizinstudenten erwähnt wird. Des Weiteren wird analysiert welche spezifischen Unterschiede zwischen Screening- und Diagnostikverfahren im verfügbaren Lehrmaterial hervorgehoben werden.

Das maßgebliche Ziel dieser Doktorarbeit ist es, eine Fragensammlung für Leser einer Screeningstudie zu entwickeln, welche die Validität und klinische Relevanz einer Screeningstudie evaluieren und messen kann. Diese Sammlung an Fragen wurde auf die bekannte Studie zum Lungenkarzinom-Screening mithilfe eines Low-Dose CTs aus dem NEJM (New England Journal of Medicine) vom Jahre 2011 [1] angewendet.

Zunächst wurde die zu untersuchende Studie ausführlich analysiert. Aufgrund meiner Arbeit als Tutor in Kursen zur Studienanalysierung, unter der Leitung von Herrn Professor Porzsolt, besaß ich die Fähigkeit wissenschaftliche Arbeiten zu bewerten. Das zur Identifikation systematischer Fehler für Therapiestudien entwickelte Instrument [66] eignete sich nicht, um die systematischen Fehler dieser Screeningstudie zu erkennen. Wissenschaftliche Instrumente, die diese Fehler entdecken sind bisher kaum verfügbar [24]. Dabei konnten durch meine Recherchen einige Ungereimtheiten in der untersuchten Studie zum Lungenkarzinom-Screening [1] aufgedeckt werden, die möglicherweise durch das Auftreten von systematischen Fehlern (Bias) bedingt waren. Diese Erkenntnisse deckten sich jedoch nicht mit den Ergebnissen aus dem USP (Usability of scientific publication)-Fragebogen, der primär für die Evaluation von Therapiestudien entwickelt wurde [66]. Resultierend kann daraus geschlossen werden, dass ein Fragebogen, der sowohl für Therapie-, Diagnostik- als auch für Screeningstudien angewendet werden kann, nicht sinnvoll erscheint. Durch eine ausführliche Literaturrecherche und zahlreiche Diskussionen in unserer Screening-Gruppe kann ich einerseits die Unterschiede zwischen Screening-, Diagnostik- und

Therapiestudien aufzeigen und andererseits die Bedeutung der wichtigen bisher publizierten Instrumente zur Analyse von Studien benennen. Durch diese Beschäftigung mit Screeningmaßnahmen und durch die bisher verfügbaren Instrumente zur Bewertung dieser, habe ich Items generiert mit welchen verschiedene Formen von Bias in Screeningstudien erkannt werden können. Diese Bias wurden in einem zweiten Schritt zu kurzen und detaillierten Fragen formuliert und zu einer Fragensammlung zusammengefasst. Diese vorläufige Fragensammlung wurde mehrmals überarbeitet und auf Verständlichkeit, Eindeutigkeit und Durchführbarkeit überprüft. Anschließend wurde diese auf die Studie aus dem NEJM angewendet und vermag diejenigen Bias aufzudecken, auf welche in dieser Studie geachtet werden sollte.

2.1 Literaturrecherche

Um herauszufinden, welche Informationen Medizinstudenten über Screeningmaßnahmen und die Bewertung dieser während ihrer Ausbildung erhalten, wurde die Approbationsordnung für Ärzte, die zum 01.01.2014 in Kraft trat, untersucht, um einen Einblick zu erhalten, welche Ziele und Erkenntnisse Mediziner im Hinblick auf die Prävention erlernen und erhalten sollen [8]. Überdies wurden alle Standardlehrwerke der Inneren Medizin in der neuesten Auflage, welche in der Bibliothek der Universität Ulm verfügbar waren studiert, um zu ermitteln wie und in welchem Umfang das Thema Prävention behandelt und dargestellt wurde. Hierbei wurde ein besonderes Augenmerk darauf gelegt, ob sowohl über die Vorteile als auch über die möglichen Nachteile von Präventions- und speziell von Screeningmaßnahmen berichtet wird. Dabei habe ich mich auf die Werke der Inneren Medizin beschränkt, da dieses Gebiet ein Hauptfach in der medizinischen Ausbildung darstellt und nahezu jeder Medizinstudent zumindest eines dieser Werke besitzt und regelmäßig in diesen studiert. Zudem nimmt das Thema Prävention ein zentrales Thema in der Inneren Medizin ein (zum Beispiel Dickdarm-Screening, Prävention von Folgekrankheiten durch den Diabetes

mellitus, Prävention von Erkrankungen des Herz-Kreislauf Systems oder Karzinomerkrankungen).

Für die Entwicklung einer Fragensammlung zur Evaluierung von Screeningstudien wurde eine Literaturrecherche über bereits existierende Instrumente und Methoden zur Beurteilung von Therapie-, Diagnostik- und Screeningstudien vorgenommen. Die Literaturrecherche wurde in vier Datenbanken, der HTA, Embase, Medline und Cochrane Library durchgeführt. Zusätzlich wurde eine freie Suche innerhalb Google/Scholar und Google/Buchsuche vorgenommen. Die durch diese Suche identifizierten Literaturwerke und dadurch entdeckten weiteren Literaturhinweisen wurde ebenfalls nachgegangen. Die Recherche erfolgte mit folgenden Suchbegriffen:
„assessing“, „quality“, „improvement“, „outcome“, „evaluation“, „quantifying value“, „efficacy“, „screening“, „screening test“, „diagnostic“, „diagnostic test“, „therapeutic“, „test“, „study“, „trial“, „reproducibility“, „impact and test“, „quality and reporting“, „complete and reporting“, „checklist“, „questionnaire“, „guidelines“. Die Suche wurde mit diesen Schlüsselbegriffen oder deren deutschsprachiger Übersetzung sowie mit verschiedenen Kombinationen aus diesen ausgeführt. Dabei wurden alle deutsch und englischsprachigen Werke verwendet, deren Volltext über die Bibliothek der Universität Ulm oder über die Fernleihe verfügbar waren. Die Ergebnisse dieser Literaturrecherche werden in Kapitel 2.2 „Analyse der Mittel zur Evaluierung von Screening-, Diagnostik- und Therapiestudien“ vorgestellt.

Um exakte Kenntnisse über den Ablauf eines Screeningverfahrens und über die notwendigen Voraussetzungen zur Implementierung eines Screenings zu erlangen, wurde eine weitere Literaturrecherche vorgenommen. Die Literatursuche wurde wieder in den bereits genannten vier Datenbanken (HTA, Embase, Medline, Cochrane Library) sowie in Google/Scholar und Google/Buchsuche ausgeführt. Dies wurde mit den folgenden Suchbegriffen und mit deren Kombinationen und gegebenenfalls mit ihrer deutschen Übersetzung vorgenommen: „screening“, „implementation“, „prevention“, „disease“, „advantage“, „disadvantage“, „positiv effects“, „negative effects“, „problems“, „impact and screening“. Zusätzlich wurden

Bücher in der Zentralbibliothek der Universität Ulm zu dem Thema „Screening“ gesucht. Die Ergebnisse dieser Literaturrecherche werden in Kapitel 2.3 „Analyse des Ablaufs eines Screeningverfahrens und Dokumentierung der Unterschiede zu diagnostischen und therapeutischen Studien“ aufgezeigt.

Für die Analyse der Studie „Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening“ wurde der Studienbericht mit genanntem Topik aus dem NEJM vom Jahre 2011 [1] und dem dazugehörigen Supplementary Appendix, der die Appendices 1-4 enthält verwendet [2]. Zudem wurde das zugehörige Studienprotokoll, das 924 Seiten aufweist und mehrere Appendices enthält [3], für die Evaluierung und überdies der Bericht „The National Lung Screening Trial: Overview and Study Design“ [4] mit den Appendices E1-E9 benutzt [5]. Ferner wurde noch der Artikel, der ebenfalls vom NLST (National Lung Screening Trial) im NEJM 2013 publiziert wurde, mit dem Titel „Results of Initial Low-Dose Computed Tomographic Screening for Lung Cancer“ [16] und der im NEJM 2014 veröffentlichte Artikel mit dem Titel „Cost-Effectiveness of CT Screening in the National Lung Screening Trial“ [12] studiert und für die Analyse verwendet. Darüber hinaus wurden, um die detailliert gestellten Fragen meiner Fragensammlung wissenschaftlich korrekt beantworten zu können, Literaturzitate aus Büchern oder den bereits mehrmals benannten Datenbanken gesucht, die die dargelegten Antworten in Kapitel 3.3 belegen.

2.2 Analyse der Mittel zur Evaluierung von Screening-, Diagnostik- und Therapiestudien

Da ständig neue Screening- und Diagnostikstudien entwickelt werden, könnte erwartet werden, dass auch die Methoden zur Überprüfung dieser Studien ständig verbessert werden. Leider existiert in diesem Bereich jedoch nur wenig verfügbare Literatur [24]. Denn fehlerhafte Studien, die Ergebnisse aufgrund von Bias (systematische Fehler) inkorrekt darstellen, können zu fehlerhaften

Schlussfolgerungen führen, Ressourcen verbrauchen und die Kosten für das Gesundheitssystem unnötig steigern [27]. Im Folgenden werden die Instrumente aufgezeigt, die häufig zitiert werden oder die sich von bereits erwähnten Verfahren in diesem Kapitel deutlich unterscheiden.

In dem Bericht „Research methods & reporting“ von CONSORT (Consolidated Standards of Reporting Trials) 2010 wurden einige Zahlen für Studien angegeben, die lückenhaft und ungenau berichtet wurden. Zum Beispiel wiesen in Pubmed im Jahre 2000 nur 21% von 519 Studien Informationen zur Verteilung der Teilnehmer auf die zu vergleichenden Gruppen auf und nur 34% von 616 im Jahre 2006. Auch die Definition eines primären Endpunktes wiesen nur 45% der Studien im Jahre 2000 und 53% der Studien im Jahre 2006 auf [48].

Aus diesen Zahlen wird die Bedeutung der Evaluation von wissenschaftlichen Arbeiten ins Bewusstsein gerückt. Zwischen 1978 und 1993 wurden 1302 Artikel in vier bedeutenden Journals (New England Journal of Medicine, Journal of the American Medical Association, British Medical Journal und Lancet) der Medline Database veröffentlicht, die sich mit diagnostischen Test Studien befassten und dabei wurde festgestellt, dass viele der diagnostischen Tests fehlerhaft beurteilt wurden aufgrund von mangelnden Informationsangaben bezüglich des untersuchten Tests [72].

Ein Artikel aus dem Jahr 1988 „A Review on the Methodology for Assessing Diagnostic Tests“ [39] wurde häufig zitiert und befasste sich mit der quantitativen Berechnung der Güte eines neuen Testverfahrens.

Mithilfe einer Vierfeldertafel können die Sensitivität, die Spezifität, die prädiktiven Werte und die Testgenauigkeit berechnet werden [39][40].

Tabelle 1: Vierfeldertafel

Mit einer Vierfeldertafel lassen sich Aussagen zur Testgenauigkeit, Sensitivität, Spezifität, positiv und negativ prädiktivem Wert treffen.

(A: Anzahl der tatsächlich Kranken, die korrekterweise ein positives Testergebnis erhalten.
B: Anzahl der tatsächlich Gesunden, die fälschlicherweise ein positives Testergebnis erhalten.
C: Anzahl der tatsächlich Kranken, die fälschlicherweise ein negatives Testergebnis erhalten.
D: Anzahl der tatsächlich Gesunden, die korrekterweise ein negatives Testergebnis erhalten.
Krank: Anzahl der Personen mit der gesuchten Erkrankung. Gesund: Anzahl der Personen ohne die gesuchte Erkrankung. Test positiv: Das Testergebnis legt die Annahme nahe, dass die untersuchte Person an der gesuchten Erkrankung leidet. Test negativ: Das Testergebnis legt die Annahme nahe, dass die untersuchte Person nicht an der gesuchten Erkrankung leidet.)

	Krank	***Gesund***	
Test positiv	**A**	**B**	**Σ A+B**
Test negativ	**C**	**D**	**Σ C+D**
	Σ A+C	**Σ B+D**	**Σ A+B+C+D**

Die Berechnung der Testgenauigkeit erfolgt durch die Bildung folgenden Quotienten:

Testgenauigkeit = (D+A) / (D+A+B+C)

Aus dieser Vierfeldertafel lässt sich auch die Sensitivität und die Spezifität eines Tests berechnen.

Sensitivität = A / (A+C)
Spezifität = D / (D+B)

Zum anderen lassen sich prädiktive Werte berechnen.
Der positiv prädiktive Wert beschreibt, wie hoch die Wahrscheinlichkeit ist bei positivem Testergebnis (Tpos) auch tatsächlich krank (K) zu sein.

P(K|Tpos) = A / (A+B)

Der negative prädiktive Wert beschreibt die Wahrscheinlichkeit bei negativem Testausgang (Tneg) auch tatsächlich gesund (G) zu sein.

$$P(G|Tneg) = D / (D+C)$$

Eine Möglichkeit der grafischen Darstellung bietet die sogenannte ROC-Kurve. Hierbei wird für jeden möglichen Cutoff-Wert des Indextests die Sensitivität und die Spezifität berechnet. Dabei wird auf die x-Achse der Wert (1-Spezifität) aufgetragen und auf die y-Achse die Sensitivität. Je weiter die entstehende Kurve von der 45 Grad Geraden entfernt ist, die durch den Nullpunkt der beiden Achsen geht und jene Punkte beschreibt, die einen nutzlosen Test kennzeichnen, umso mehr eignet sich der Test, ein vermutetes Ergebnis zu bestätigen oder auszuschließen [98].

1996 wurde durch die erstmalige Einführung einer Leitlinie, dem CONSORT Statement („Consolidated Statement of Reporting Trials“) für randomisierte kontrollierte Therapiestudien ein erheblicher Fortschritt erzielt. Diese Leitlinie wurde bis heute zweimal überarbeitet: 2001 und 2010. „Das CONSORT 2010 Statement gibt Autoren Empfehlungen für die Erstellung von Berichten von randomisierten kontrollierten Studien (...)“ [83, S. e20]. An dieser Stelle soll betont werden, dass diese Leitlinie primär für die Ersteller zur korrekten Berichterstattung von Studien konzipiert wurde, nicht aber für die Bewertung der Qualität einer Studie. Die Leitlinie beinhaltet ein Flussdiagramm und eine Checkliste mit 25 Punkten [83]. Mit dem CONSORT Flussdiagramm soll, um die notwendige Transparenz herzustellen, der Verlauf von der Aufnahme der Studienteilnehmer in die Studie bis zur finalen Analyse der Studienergebnisse auf einen Blick ersichtlich sein.

Die CONSORT Checkliste (Tabelle 2) und das CONSORT Flussdiagramm (Abbildung 1) sind im Anhang dieser Arbeit abgebildet. Die CONSORT Checkliste gibt Autoren und Editoren einen Wegweiser, um eine korrekte und vollständige Veröffentlichung von Therapiestudien zu ermöglichen, bei denen auch das Patienten-Outcome im Mittelpunkt steht [23]. Zusätzlich ist noch eine 28 Seiten lange Erklärung zu dem Umgang mit den Leitlinien und der Handhabung der

Checkliste und des Flussdiagramms online einsehbar [48]. Auf einer eigenen Homepage unter www.consort-statement.org wird über die Entwicklung und mehrmalige Überarbeitung der Leitlinie ausführlich berichtet.

CONSORT gilt als einer der meist zitierten Beiträge, der weltweit eine Anzahl von über 5300 Zitaten (selbst-Zitate nicht eingeschlossen) aufweist. Es wurden auch systematische Reviews angefertigt, um über den Benefit von CONSORT zu berichten. Ein Cochrane Review von 2012 beschrieb den relativen Nutzen von CONSORT bei 69 von 81 Metaanalysen (1996-2001) im Zusammenhang mit der lückenlosen Dokumentation von Studien. Dennoch bleibt schlussfolgernd die vollständige Berichterstattung suboptimal [94].

Ein weiteres systematisches Review von 2006 beschäftigte sich mit der Qualitätsverbesserung von Studienberichten durch die CONSORT Checkliste. Verglichen wurden Studien im Zeitraum von 1996 bis 2005, welche die Checkliste nutzten mit denen, die ohne Checkliste von CONSORT arbeiteten. Die Qualität der Berichterstattungen, die mit der CONSORT Checkliste erstellt wurden, war signifikant besser als die Qualität der Berichte ohne Anwendung von CONSORT [63]. Die Geschichte zur Entwicklung des CONSORT Statements ist bedeutend, um die spätere Entwicklung von Instrumenten zur Bewertung von Diagnostik- und Screeningverfahren zu verstehen.

Ein weiteres bedeutendes Paper, das die Methodik zur Güteprüfung einer Studie veränderte, war „The STARD Initiative“, die im Jahre 2003 erschienen war [13]. Die Abkürzung „STARD“ steht für: Standards for Reporting of Diagnostic Accuracy. Dieses Paper ist eine wissenschaftliche Abhandlung, das aus einer internationalen und multidisziplinären Zusammenarbeit von Forschern aus verschiedenen Disziplinen der Wissenschaft und Medizin, sowie Redakteuren von führenden Journals, entstanden war. Dieses Werk verfolgt die Absicht, Autoren eine Leitlinie zu geben, mit der sie die Durchführung von Diagnostikstudien evaluieren können [23]. Es ist hervorzuheben, dass die „STARD Initiative“ als eines der wenigen Werke auch die Bewertung von Screeningtests miteinbezieht. Das STARD Komitee erstellte in Anlehnung an das CONSORT bei Ihrem gemeinsamen Treffen einerseits eine 25 Punkte umfassende Checkliste und eine Anleitung für die korrekte Dokumentation von Studien, andererseits fertigte sie ein Flussdiagramm

an, das Informationen zur Patientenrekrutierung, Ausführung des Tests und die Anzahl an Patienten, die den Index- und Referenztest durchlaufen angibt. Der Indextest ist jener Test, dessen Nutzen und Überlegenheit in einer Studie gezeigt werden soll. Verglichen wird dieser Test mit dem Referenztest, der den momentan besten verfügbaren Test und meist den aktuellen Goldstandard darstellt [74].

Die STARD Initiative versucht mithilfe dieser zwei Mittel die Dokumentation von Studien zu verbessern und Autoren zu helfen, bei der Berichterstattung elementare Schritte nicht zu übersehen. Die 2003 veröffentlichte Checkliste (Tabelle 3) und das Flussdiagramm (Abbildung 2) sind im Anhang der Arbeit abgebildet.

Das STARD Komitee hat sich darauf geeinigt, ausschließlich ein Verfahren zu entwickeln, mit dessen Hilfe die Qualität eines Screening- oder Diagnostiktests gemessen werden kann. Hingegen wurden die mittel- oder langfristigen Erfolge für das Patienten-Outcome nicht berücksichtigt [23]. Bis Februar 2013 wurde STARD in über 200 biomedizinischen Journalen aufgenommen und als Handlungsauftrag für ihre Autoren übernommen. Des Weiteren wurde das STARD Statement über 1600 mal zitiert und wurde seit der Ersterscheinung 2003 noch nie überarbeitet. Doch zehn Jahre nach der Publikation lässt sich nur ein langsamer Fortschritt der Qualitätsverbesserung über die Dokumentation diagnostischer Studien verzeichnen [53], wie es unter anderem auch die Studie von Selman und Kollegen 2011 [84] zeigte.

In demselben Jahr 2003 erschien ein weiterer Artikel, der auf die STARD Studie Bezug nahm und ein Instrument vorstellte mit dem Diagnostikstudien bewertet werden können. Hierbei handelte es sich um QUADAS (Quality Assessment of Diagnostic Accuracy Studies), welches eine Checkliste mit 14 Themen umfasst und von neun Experten entwickelt wurde [101]. Diese Themen wurden zu Fragen formuliert und sind in der Tabelle 4 im Anhang der Arbeit dargestellt. Diese Methode zur Qualitätsprüfung wurde ausschließlich für diagnostische Studien generiert.

Im August 2006 wurde im AJR (American Journal of Roentgenology) eine weitere bedeutende Metaanalyse [24] für die Evaluation von Testgenauigkeit bei

Diagnostik- und Screeningstudien publiziert. Drei wichtige Aspekte wurden in diesem Paper hervorgehoben:

1. Durchführung des Tests und Bestimmung der Testgenauigkeit und die prädiktiven Werte für den Test
2. Bedeutung des Tests im Hinblick auf die weiterführende Behandlung
3. Bedeutung des Tests für das Outcome des Patienten

Zudem wurden sechs wichtige Stufen benannt, die für das Erreichen einer exakten Dokumentation einer Studie elementar sind:

1. Definition der Studienziele
2. Geeignete Literaturrecherche
3. Überprüfung der Studienqualität und Durchführbarkeit
4. Darlegung der Datenherkunft
5. Statistische Analyse
6. Ergebnisinterpretation und Entwicklung von Verbesserungsvorschlägen

Im Weiteren wurde noch auf statistische Methoden hingewiesen und diese erläutert, ähnlich wie im Paper von 1988 „A Review on the Methodology for Assessing Diagnostic Tests“.

2009 entstand das PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) Statement mit einer 29 Items umfassenden Checkliste, welche die Berichterstattung von systematischen Reviews und Metaanalysen verbessern sollte und oft zitiert wurde [68]. Die Checkliste unterschied sich jedoch nicht wesentlich von bereits früher publizierten Werken von CONSORT und STARD.

2012 entstand CASP (Critical Appraisal Skills Programm) unter der Leitung von Sir Muir Gray [17]. Es wurden mehrere Checklisten für verschiedene Studientypen publiziert. Die Checkliste für diagnostische Studien ließ keine große Differenz zum USP-Fragebogen (CASP beinhaltet zusätzlich Fragen zur Randomisation der Teilnehmer und zur Abschätzung der Studienergebnisse) [66], der auf die Studie aus dem NEJM angewendet wurde, erkennen, weshalb ich auf CASP nicht näher

eingehen werde.

Ein oft zitiertes Review wurde 2012 unter dem Titel „Quantifying the Accuracy of a Diagnostic Test or Marker“ veröffentlicht [40]. In dieser wissenschaftlichen Abhandlung lag der Fokus auf der Untersuchung der klinischen Validität und der klinischen Anwendbarkeit und des Nutzens eines Tests. Der klinische Nutzen beschreibt, um wie viel Grad ein neuer Test (Indextest) zu einem verbesserten therapeutischen Management und zu einem verbesserten Patienten-Outcome führt. Denn nicht jeder Indextest, der dem Referenztest überlegen ist, führt auch folglich zu einem erfolgreicheren therapeutischen Verfahren. Ein Schaubild aus dem Review fasst diese Einteilung in analytische Validität, klinischer Nutzen und klinische Validität zusammen. Diese Abbildung ist im Anhang (Abbildung 3) verzeichnet. In diesem Review wurde die Bedeutung der sogenannten „single-test accuracy study“ hervorgehoben. Darunter versteht man, dass sowohl der zu untersuchende Test (Indextest), als auch der Referenztest auf ein und dasselbe Individuum angewendet wird, von dem vermutet wird, Träger der gesuchten Erkrankung zu sein. Das Individuum bekommt unabhängig vom Ergebnis des Indextests auch die Untersuchung mit dem Referenztest und ist gegenüber dem Ergebnis des Indextests verblindet, das heißt dem betroffenen Individuum wird das Testergebnis nicht mitgeteilt. Dieses Verfahren besaß die höchste Aussagekraft, wenn es sich bei dem Studiendesign um eine prospektive, kontrollierte, randomisierte Studie handelte. In diesem Paper wurde auch ein Augenmerk darauf gelegt, dass das Krankheitsstadium, in dem sich ein Individuum befindet von ausschlaggebender Wichtigkeit für die spätere Anwendbarkeit eines Tests ist. Überdies konnte der Prototyp der single-test accuracy study nicht mit jedem Test durchgeführt werden. Stellte der Referenztest eine Biopsieentnahme beispielsweise dar, konnte dies nicht bei jedem Studienteilnehmer, aus ethischen Gesichtspunkten, durchgeführt werden. Bei dieser Art der Studie wurden dann nur diejenigen Patienten zur Biopsie (Referenztest) geladen, bei denen der Indextest positiv war. Dies bedeutete wiederum, dass in diesem Fall nur die Anzahl der falsch Positiven herausgefunden wurde, nicht aber die Anzahl der falsch Negativen, sodass schlussfolgernd nur der positiv prädiktive Wert berechnet werden konnte. Dies nennt man in der Fachsprache eine differenzielle Verifikation

[40].

Ransohoff und Feinstein [71] publizierten 1978 im NEJM ein Paper mit dem Titel „Problems of spectrum and bias in evaluating the efficacy of diagnostic tests“. Die beiden Autoren fanden heraus, dass vielversprechende Tests mit guten Ergebnissen, die bei Patienten mit einem fortgeschrittenen Erkrankungsstadium durchgeführt wurden, ihre Aussagekraft oft bei Patienten, die in einem Anfangsstadium waren, verloren. War die Prävalenz einer Erkrankung (Relative Anzahl der Menschen in einem bestimmten Kollektiv zu einem bestimmten Zeitpunkt mit der gesuchten Erkrankung) hoch, so würde in diesem Patientenkollektiv die Erkrankung mit fortgeschrittenem Stadium überwiegen (Krankenhaus) und umgekehrt. Aus diesem Grund sollte beachtet werden, dass die Studienpopulation mit der Zielpopulation, auf die der Test später angewendet werden soll, vergleichbar ist.

Aus der Zusammenschau der aufgelisteten Mittel in diesem Kapitel wird ersichtlich, dass bereits einige analytische sowie auch quantitative Mittel zur Evaluierung von Tests beschrieben wurden. Heutzutage existiert jedoch, wie es auch aus der Literaturrecherche hervorgeht, insgesamt nur wenig Literatur zu diesem Thema. Diese Problematik wurde auch in der Metaanalyse von 2006 [24], wie bereits in diesem Kapitel beschrieben, erwähnt. Zudem wird erst in den neueren wissenschaftlichen Veröffentlichungen der letzten Jahre die Bewertung von Screeningtests beachtet. Allerdings fehlen adäquate Studien und Paper, die sich mit der Analyse und Evaluierung von Screeningstudien gesondert beschäftigen. In all den aufgeführten Werken werden, falls Screeningtests überhaupt berücksichtigt werden, nur gemeinsame Richtlinien, Checklisten und Grafiken mit Diagnostikstudien angeboten. Doch ob diese Handhabung sinnvoll ist und welche Unterschiede es zwischen Diagnostik-, Therapie- und Screeningstudien gibt, werde ich in dem folgenden Kapitel 2.3 erörtern. Ebenfalls werde ich in der Diskussion die Unterschiede und Parallelen der hier erwähnten Methoden zu meinem Vorschlag für die Beurteilung von Screeningstudien aufzeigen.

2.3 Analyse des Ablaufs eines Screeningverfahrens und Dokumentierung der Unterschiede zu diagnostischen und therapeutischen Studien

2.3.1 Unterschiede zwischen den drei Arten der Prävention

Um Screeningstudien bewerten und evaluieren zu können, muss ein Grundverständnis für den allgemeinen Ablauf eines Screeningverfahrens geschaffen werden. In meiner Promotionsarbeit beschränke ich mich auf die Analyse von Studien der sekundären Prävention und dabei speziell auf Screeningstudien. Die Unterschiede zwischen den drei Arten der Prävention zeigt die Tabelle 5.

Tabelle 5: Unterschiede zwischen der primären, sekundären und tertiären Prävention (abgewandelte Version aus der Arbeit „Begriffshygiene“ von Porzsolt F, 2014 [65])

Die folgende Darstellung verdeutlicht, dass jede Präventionsart (primär, sekundär oder tertiär) unterschiedliche Teilschritte durchläuft um einen klinisch relevanten Erfolg zu generieren. Die farbig unterlegten Kästchen stellen die Zugehörigkeit zu der jeweiligen Präventionsart dar. Die weißen Kästchen bedeuten, dass diese Maßnahmen hier keine Anwendung finden.
(Primäre Prävention: Das Auftreten einer Krankheit soll verhindert werden. Sekundäre Prävention: Eine Krankheit soll frühzeitig diagnostiziert werden. Tertiäre Prävention: Folgeschäden oder Rezidivereignisse einer Krankheit sollen frühzeitig diagnostiziert werden.)

	Screening	*Diagnostik*	*Therapie*
Primäre Prävention			
Sekundäre Prävention			
Tertiäre Prävention			

Aus der Tabelle 5 lassen sich bereits wichtige Schritte erkennen, die eine Screeningmaßnahme durchläuft. Die sekundäre Prävention gliedert sich in die Hauptschritte: Screening, Diagnostik und Therapie. Maßnahmen zur primären Prävention beschränken sich darauf, durch gezielte Verhaltensänderungen Krankheiten zu vermeiden (Therapie) [103]. Bei der tertiären Prävention hingegen wird versucht, durch eine regelmäßige Diagnostik Rezidivereignisse oder

Folgeschäden einer Krankheit frühzeitig zu erkennen und anschließend spezifische therapeutische Maßnahmen einzuleiten (Diagnostik und Therapie) [54].

2.3.2 Voraussetzungen zur Implementierung eines Screenings

In verschiedenen wissenschaftlichen Artikeln werden Richtlinien angesprochen, die beachtet werden sollten, um die erfolgreiche Implementierung von Screeningprogrammen zu ermöglichen [78][34][49]. Denn trotz der enormen Popularität, die das Screening für Krankheiten heutzutage erfährt, sollte immer eine Abwägung zwischen dem erwarteten Nutzen für die Patienten und dem möglichen Schaden, welcher durch die Verbreitung von Angst, Ungewissheit und durch den Verbrauch von medizinischen Ressourcen, getroffen werden [34][49]. Um eine Erkrankung in einem frühen Stadium mittels Screening entdecken zu können, muss das symptomfreie Intervall ausreichend lange sein. Außerdem sollte zu diesem Zeitpunkt die gesuchte Erkrankung (siehe rote Markierung in der Abbildung 4) noch erfolgreich therapierbar sein [70].

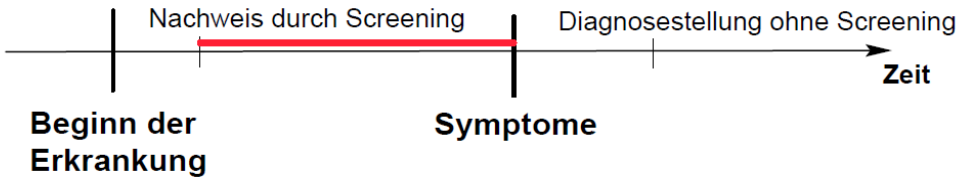

Abbildung 4: Zeitabfolge einer Erkrankung

In der obigen Darstellung wird auf einer Zeitachse gezeigt, in welcher zeitlichen Abfolge das Screening in Bezug zu Beginn und Symptomen einer Erkrankung steht.

(Beginn der Erkrankung: Ab diesem Zeitpunkt könnte die Erkrankung theoretisch nachgewiesen werden. Nachweis durch Screening: Ab diesem Zeitpunkt ist eine Diagnosestellung mittels Screening möglich. Symptome: Zu diesem Zeitpunkt treten die ersten Symptome der Erkrankung auf. Diagnosestellung ohne Screening: Zu diesem Zeitpunkt erfolgt ohne Screening die Diagnose der Erkrankung. Rote Linie: Intervall, indem ein Screening stattfinden sollte.)

Walter Holland verfasste 1993 einen Artikel über die Implementierung eines Screenings [34], in dem er über drei entscheidende Grundlagen berichtete, die vor der Einführung eines Screenings geprüft werden sollten. Erstens wird ein vernünftig geführtes Bevölkerungsregister für ein Screening benötigt, aus welchem die Zielgruppe, die dem Screening zugeführt werden soll, ermittelt wird. Zweitens müssen die Institutionen, die ein Screening durchführen, über die notwendigen Mittel, über ausreichend medizinische Qualität und über das medizinische Wissen verfügen, um eine adäquate Diagnostik, Therapie und Folgemaßnahmen einleiten zu können. Denn nur wenn diese Punkte gewährleistet sind, kann ein Nutzen für die Patienten erreicht werden. Drittens sollte eine national tätige Institution eingeführt werden, die das Screening überwacht, evaluiert und neue empfohlene Tests vor deren Einführung überprüft.

Rose und Barker setzten in ihrer Arbeit „Screening“ [78] ebenfalls auf drei Sachverhalte, die vor der Etablierung eines Screenings beachtet werden sollten. Zum einen sollte das Ergebnis eines Screenings anhand des Effektes auf die Mortalität und Morbidität der Patienten gemessen werden. Zum anderen sollte die Validität und Reliabilität eines Screeningtests ausreichend hoch sein und dies anhand von Sensitivität, Spezifität und prädiktiven Werten gemessen werden und zudem sollte nachwiesen werden, dass die entdeckten Krankheitsfälle eine bessere Prognose durch die frühzeitige Entdeckung aufweisen.

Ein weiterer Punkt, den Wilson und Jungner zu den bereits genannten Aspekten noch hinzufügten ist, dass durch multiples Screening andere wichtige Aspekte der medizinischen Versorgung vernachlässigt werden. Einerseits durch eine differente Ressourcenverteilung zu Gunsten des Screenings und andererseits durch die Induktion einer falschen Sicherheit, die durch das Screening propagiert wird [103].

2.3.3 Screeningmodelle

Zunächst betrachten wir ein ideales Screeningmodell, siehe auch Abbildung 5, welches schnell, billig und sicher ist. Alle durchgeführten Tests zeigen immer ein positives oder negatives Ergebnis an [99]. Für den Ablauf eines idealen Screeningprogramms werden aus einer Gruppe von Personen mit vermutetem Risikoprofil für die gesuchte Erkrankung, die Personen mit dem höchstem Risiko zur sogenannten Hochrisikogruppe zusammengefasst. Nur die Hochrisikogruppe wird aus folgenden Gründen zum Screening zugelassen:

1. Bei geringer Prävalenz der Krankheit nimmt im untersuchten Patientenkollektiv (das heißt der Anteil der Erkrankten im Kollektiv), der positiv prädiktive Wert ab, das heißt „(...) die Wahrscheinlichkeit (...), bei positivem Testausgang krank zu sein“, sinkt [64, S. 45].
2. Die gescreenten Personen werden einer psychischen Belastung, möglichen Komplikationen und unangenehmen Prozeduren ausgesetzt, die anhand der oft geringen Prävalenz der gesuchten Erkrankung gegenüber den betroffenen Patienten vertretbar und verantwortbar sein muss. Diese Belastung der Betroffenen zeigte die Studie von Thornton H. 2003, welche die möglichen negativen Folgen für Frauen beim Mammographie-Screening verdeutlichte [91].
3. Aus ökonomischen Gesichtspunkten und zur Vermeidung eines unnötigen Ressourcenverbrauchs macht es Sinn, ein Screeningprogramm für eine Personengruppe anzubieten, die ein erhöhtes Risiko trägt an der gesuchten Erkrankung zu leiden.

Bei einem positiven Ausgang des Screeningtests werden den betroffenen Personen aus der Hochrisikogruppe diagnostische Tests zur Ergebnisverifikation empfohlen, um gegebenenfalls bei positivem Ausgang eine geeignete Therapie einleiten zu können. Bei negativem Testausgang, sowohl beim Screeningtest als auch bei den diagnostischen Untersuchungen, werden die Personen nach einer bestimmten Zeit erneut zum Screening eingeladen.

In der folgenden Abbildung 5 wird die aufgearbeitete Thematik dargestellt.

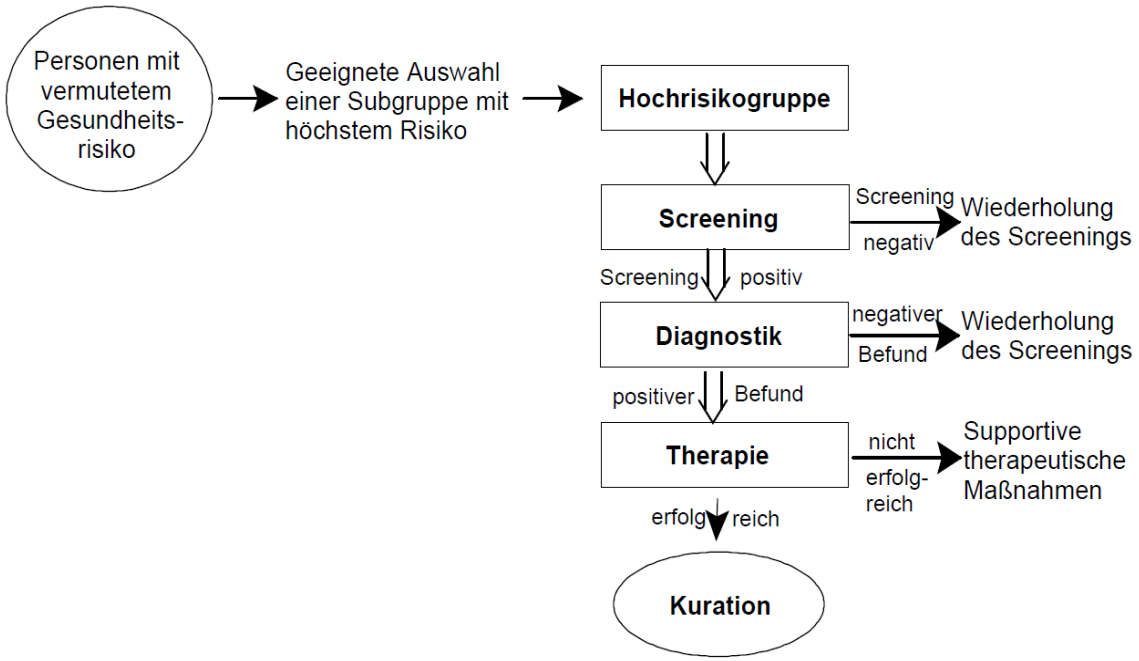

Abbildung 5: Ablauf eines idealen Screeningprogramms

In der obigen Darstellung werden die einzelnen Teilschritte der sekundären Prävention illustriert, die durchlaufen werden müssen von der Rekrutierung einer Hochrisikogruppe bis zur Therapie der gesuchten Erkrankung. Bei den ersten beiden Teilschritten nach der Auswahl der Hochrisikogruppe „Screening“ und „Diagnostik“ kann das Untersuchungsergebnis jeweils dichotom ausfallen. Bei einem positiven Ergebnis wird gemäß der oben dargestellten Reihenfolge ein neuer Teilschritt bis zur Therapie durchlaufen. Bei einem negativen Ergebnis beginnen die betroffenen Personen wieder vor dem ersten Teilschritt, dem Screening.

(Screening positiv: Das Testergebnis legt die Annahme nahe, dass die untersuchte Person an der gesuchten Erkrankung leidet. Screening negativ: Das Testergebnis legt die Annahme nahe, dass die untersuchte Person nicht an der gesuchten Erkrankung leidet. Positiver Befund: Die gesuchte Erkrankung wurde nachgewiesen. Negativer Befund: Die gesuchte Erkrankung wurde nicht nachgewiesen. Wiederholung des Screenings: Die Screeninguntersuchung wird den Personen nach einer bestimmten Zeitspanne erneut empfohlen (zum Beispiel beim Mammographie-Screening alle zwei Jahre).)

Nun betrachten wir ein realistischeres Bild von einem Screeningprogramm. Die Testergebnisse sind nicht immer dichotom, sondern können sich auch in einer Grauzone, dem sogenannten Indifferenzbereich, zwischen positiv und negativ befinden. Entweder werden die Betroffenen erneut gescreent, einem anderen diagnostischen Test unterzogen, der weniger invasiv und komplikationsreich ist als der Goldstandard oder sie werden direkt dem Goldstandard zugeführt. Der Goldstandard ist die Untersuchungsmethode, die der Wahrheit am Nächsten kommt, jedoch ist sie nur ein Näherungsverfahren und bietet deshalb, so wie jeder andere Test, keine vollkommene Sicherheit [64]. Auch bei diesen Tests besteht die Möglichkeit, dass die gesuchte Erkrankung weder nachgewiesen noch ausgeschlossen werden kann. Diese Patienten leben für eine lange Zeit in der Ungewissheit und müssen sich vielen unangenehmen, schmerzhaften und mit Komplikationen behafteten Untersuchungen unterziehen, obschon sie möglicherweise nicht an der gesuchten Erkrankung leiden [99]. In der Abbildung 6 wird diese Problematik illustriert.

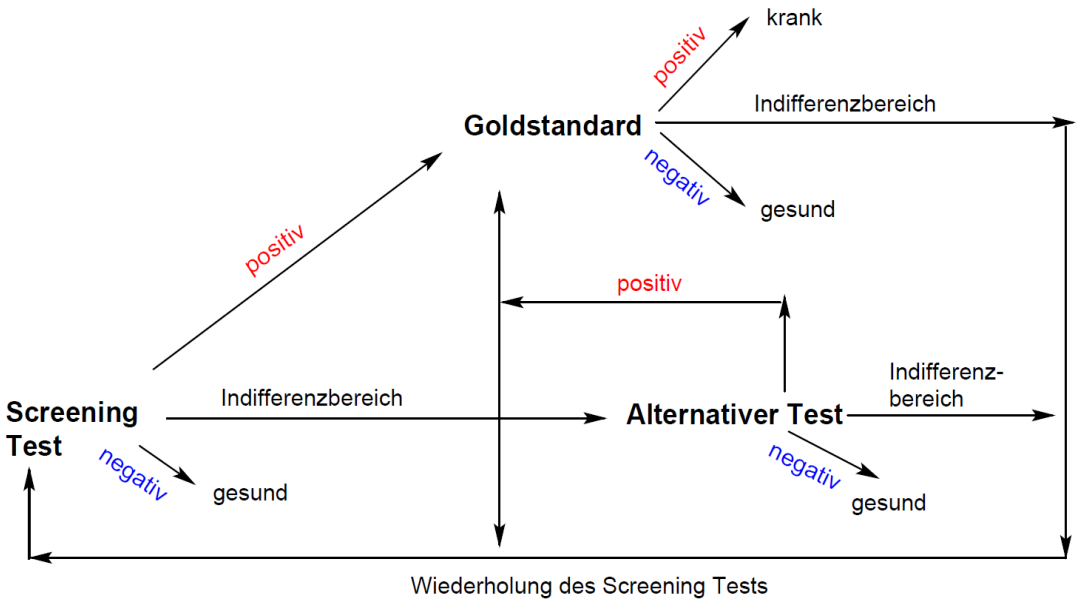

Abbildung 6: Ablauf eines realistischen Screeningprogramms (abgewandelte Version aus der Arbeit „Should i be tested for cancer?“ von Welch HG, 2004 [99])

In der obigen Darstellung soll gezeigt werden, dass jeder Test (Screening Test, Alternativer Test, Goldstandard) einen Bereich aufweisen kann, der sich weder einem positiven noch einem negativen Ergebnis zuteilen lässt (Indifferenzbereich). Für dieses betroffene Klientel gibt es beim Screening Test drei Möglichkeiten (Wiederholung des Screenings, Durchführung eines Alternativen Tests oder des Goldstandards). Welches dieser drei Verfahren angewendet wird kann der Untersucher nach individuellen Gesichtspunkten entscheiden. Ein Ergebnis im Indifferenzbereich beim Alternativen Test oder Goldstandard führt zu einer Wiederholung des Screening Tests nach einer bestimmten Zeitspanne. Ein positives Ergebnis beim Screening oder Alternativen Test führt zur Anwendung des Goldstandards. Erst ein positives Ergebnis im Goldstandard deklariert den Patienten als „krank“, indem Sinne an der gesuchten Erkrankung zu leiden. Ein negatives Ergebnis bedeutet bei allen drei Tests, dass der Patient „gesund“ ist, indem Sinne nicht an der gesuchten Erkrankung zu leiden.

(Positiv: Das Testergebnis legt die Annahme nahe, dass die untersuchte Person an der gesuchten Erkrankung leidet. Negativ: Das Testergebnis legt die Annahme nahe, dass die untersuchte Person nicht an der gesuchten Erkrankung leidet. Indifferenzbereich: Der Test kann die Erkrankung weder nachweisen noch ausschließen. Goldstandard: Der beste aktuell verfügbare Test, um die Erkrankung nachzuweisen. Alternativer Test: Ein anderer Test neben dem Goldstandard, der weniger invasiv, aber ungenauer eine Erkrankung nachweisen kann. Gesund: Die Patienten leiden nicht an der gesuchten Erkrankung. Krank: Die Patienten leiden an der gesuchten Erkrankung.)

2.3.4 Unterschiede zwischen Studien der sekundären Prävention zu diagnostischen oder therapeutischen Studien

In vielen Arbeiten, die in Kapitel 2.2 besprochen wurden, werden die Methoden zur Evaluierung von Studien entweder nur auf diagnostische Tests angewendet oder es wurden Richtlinien für diagnostische Tests und Screeningtests gemeinsam vorgestellt. Ich bin der Überzeugung, dass eine adäquate Studienanalyse mit einem Instrument für beide Studientypen nicht sinnvoll ist, da eine Studie über eine Screeningmaßnahme Fehlerquellen bei den folgenden Punkten enthalten kann:

- bei der Auswahl einer geeigneten Hochrisikogruppe
- bei der Wahl eines geeigneten Endpunktes
- bei der Screeningmaßnahme selbst
  - Compliance mit dem Screeningprogramm (regelmäßige Teilnahme am Screening)
  - Likelihood Ratio des Screenings (das Screeningprogramm muss in der Lage sein, die gesuchte Erkrankung nachzuweisen)
  - die Bewertung der Likelihood Ratio des Screeningtests
  - Ausmaß der subjektiven Beurteilung durch den Untersucher
- bei der diagnostischen Maßnahme
- bei der therapeutischen Maßnahme

Eine detaillierte Ausführung und Ergänzungen zu diesen Punkten folgt im Kapitel 3.2. Aus dieser kurzen Auflistung wird ersichtlich, dass eine Screeningstudie wesentlich mehr Fehlerquellen beinhalten kann, als dies bei therapeutischen oder diagnostischen Studien der Fall ist. Die diagnostischen Studien können Bias in den Bereichen der diagnostischen und folgenden therapeutischen Maßnahmen aufweisen. Eine Screeningstudie jedoch beinhaltet die Fehlerquellen einer diagnostischen und einer therapeutischen Studie und weist zusätzlich noch Biasformen bei der Auswahl der Hochrisikogruppe, bei der Wahl eines geeigneten Endpunktes und bei der Screeningmaßnahme selbst auf. Aus diesen Gründen

werden Evaluationsmethoden, die für beide Studientypen entwickelt werden, diesem Sachverhalt nicht gerecht. Die beiden Studientypen müssen unterschieden und getrennt voneinander mit Richtlinien und Hilfsmitteln, die für denjenigen Studientyp maßgeschneidert sind, analysiert werden, um eine adäquate Prüfung der Studien zu erreichen. Unterstützt wird meine Theorie dadurch, dass bis heute Mängel in der adäquaten Berichterstattung von Studien existieren [48] und folglich die Kontrollgremien in ihrer Arbeit auf kein standardisiertes und erfolgversprechendes Verfahren zurückgreifen können. Dies führt zu mangelhaften Berichten in der wissenschaftlichen Literatur und dadurch zur Begünstigung von Fehlentscheidungen.

2.4 Entwicklung einer Fragensammlung zur Analyse von Screeningstudien

Die Studie aus dem NEJM wurde für die Analyse ausgewählt, da sie zum Zeitpunkt des Beginns meiner Doktorarbeit die zuletzt publizierte Screeningstudie war, die für großes Aufsehen in der Medizin, vor allem unter den Radiologen, Pulmonologen und Chirurgen, sorgte. Die Studie wurde zunächst durch die Erfahrungen, die ich in den Seminaren zur Studienanalyse von Herrn Professor Porzsolt gewonnen hatte, genau untersucht und dabei Stellen markiert, die einer detaillierteren Begutachtung bedürfen. Dabei wurden erste Zweifel in Bezug auf die interne und externe Validität der Studie erhoben. Im Anschluss daran wurde der USP-Fragebogen (Usability of scientific publication-Fragebogen) [66] auf die Studie angewendet (das genaue Ergebnis des USP-Fragebogens ist im Anhang unter Tabelle 6 verzeichnet), welcher entwickelt wurde, um therapeutische Studien in Hinblick auf ihre Validität und klinische Relevanz zu überprüfen. Laut dem Ergebnis aus dem USP-Fragebogen ist die Studie mit sieben Punkten eindeutig valide. Jedoch beinhaltet dieser Fragebogen fast keine Items, die für einen Screeningprozess entscheidend sind. In der Arbeitsgruppe, die sich mit unterschiedlichen Screeningstudien beschäftigte, sind verschiedenste Probleme

bei der Analyse dieser Studien mit dem USP-Fragebogen aufgefallen. Aus diesem Grund wurde der Fragebogen erweitert und ist im Anhang in Tabelle 7 abgebildet. Aus den Erfahrungen wie ein Screeningprozess funktioniert, welches in Kapitel 2.3 beschrieben wurde und durch meine Anmerkungen in der Erstanalyse der Studie, konnte jedem Teilschritt von der Patientenrekrutierung bis zur finalen Analyse der Studienergebnisse eine passende Biasform zugeordnet werden. Einige der Biasformen wurden dann in einem zweiten Schritt zusammengefasst, wie beispielsweise der Incentive Bias und der Volunteer Bias und andere wiederum aus Gründen der Übersichtlichkeit gänzlich weggelassen wie beispielsweise der Information Bias oder der Observer Bias. Die verbliebenen Bias wurden dann den bereits beschriebenen drei großen Teilschritten (Screening, Diagnostik und Therapie) zugeordnet und als direkte Bias bezeichnet. Aus Gründen der genauen Veranschaulichung wurden alle dokumentierten Biasformen im Ergebnisteil 3.2 erläutert und definiert. Zusätzlich zu diesen Biasformen müssen noch weitere Items beachtet werden, die zu den indirekten Bias zusammengefasst wurden, um eine Studie vollständig evaluieren zu können. Diese Items entstammen dem erweiterten USP-Fragebogen. Um dem Leser einer Screeningstudie ein Instrument zu geben, das einfach anzuwenden ist, wurde jeder Biasform eine oder mehrere explizit gestellte Fragen zugeteilt.

3. Ergebnisse

3.1 Screening als Teil des Ausbildungscurriculums bei Medizinstudenten

Screeningprogramme werden in unserer Zeit in vielen medizinischen Gebieten angeboten, wie zum Beispiel in der Gynäkologie das Mammographie-Screening oder in der Urologie das PSA-Screening. Während des Studiums und in den damit verbundenen Vorlesungen an der Universität Ulm kann ich aus meiner persönlichen Erfahrung berichten, dass zu passenden Themen die Möglichkeit ein Screening durchzuführen erläutert wurde und bei bestimmten Risikofaktoren auch auf die Notwendigkeit zu screenen hingewiesen wurde. Allerdings wurde zu keiner Zeit, weder auf die number needed to screen (NNS), die Notwendigkeit eine Risikogruppe zu definieren, noch auf die falsch positiven und falsch negativen Ergebnisse, die bei einem Test auftreten, verwiesen. Dies stellt aus meiner Sicht aber einen wichtigen Aspekt dar, um im späteren ärztlichen Beruf eine Screeningmaßnahme bewerten und gegebenenfalls erfolgreich durchführen zu können. Die praktizierenden Ärzte sollten über die Anzahl der falsch positiven und falsch negativen Ergebnisse Bescheid wissen, um daraus die richtigen Handlungsschlüsse und bei bestehendem klinischen Verdacht die notwendige weiterführende Diagnostik in Erwägung zu ziehen. Außerdem, „(...) die für das ärztliche Handeln erforderlichen allgemeinen Kenntnisse, Fähigkeiten und Fertigkeiten in Diagnostik, Therapie, Gesundheitsförderung, Prävention und Rehabilitation, (...)“ beherrschen [8, S. 5]. Das Thema Prävention und Gesundheitsförderung wird in einem Querschnittsfach im klinischen Abschnitt der medizinischen Ausbildung angeboten und gelehrt [8]. Um einen Überblick darüber zu erhalten, welchen Inhalt Studenten in Büchern über Prävention lernen, wurden die großen Werke der Inneren Medizin studiert. Bedeutend ist, dass einige der großen Standardwerke der Inneren Medizin zu einem großen Teil gar nicht oder nur ganz am Rande auf die Prävention und das Screening eingehen [73][62][25] [37]. Es wird über die Diagnosestellung bei unspezifischen Symptomen und

Laborbefunden gesprochen. Über Screeningprogramme wird nur sehr oberflächlich diskutiert, man spricht in diesem Zusammenhang von „(...) indiskriminierte, d.h. ungezielte Untersuchungen im Rahmen sog. Screeningprogramme mit bestimmter Zielfunktion (...)“ [88, S. 7] und es wird darauf hingewiesen, dass der Patient über die Maßnahmen durch den behandelten Arzt hinreichend genau informiert werden sollte [88]. In einem anderen Werk wird speziell auf die Aussagekraft von Tests in Bevölkerungen mit unterschiedlicher Prävalenz einer Erkrankung hingewiesen, denn ein hervorragender Test in einer Population mit hoher Prävalenz würde in einer vergleichbaren Population mit geringer Prävalenz schlechter abschneiden und viele falsch positive Ergebnisse liefern [76]. In der neuesten Auflage dieses Buches [30] findet sich in dem Kapitel „Klinische Epidemiologie“ jedoch keine vergleichbare Aussage mehr. Dafür wird der Prävention nun ein eigenes dreiseitiges Kapitel gewidmet, in dem die drei verschiedenen Präventionsarten und die Bedingungen für die Einführung eines Screenings kurz erläutert werden [82]. Bei der sekundären Prävention wird darauf eingegangen, dass die Screeningmethode eine ausreichend hohe Sensitivität und eine hohe Spezifität braucht, „(...) um möglichst wenige falsch-positive Befunde zu erbringen.“ [82, S.203]. Es ist also von entscheidender Bedeutung in welchem Rahmen und für welches Klientel ein Test angewendet wird und aus diesen Gründen sollte eine Hochrisikogruppe ausgewählt werden, die für den Test geeignet ist. Es wird auch in einem knapp bemessenen Rahmen auf die Bewertung diagnostischer Untersuchungen und die Berechnung der Wertigkeit dieser eingegangen [30]. In einem neuen und modernen Werk der Inneren Medizin (Duale Reihe – Innere Medizin) existiert kein gesondertes Kapitel, in welchem über Prävention und Screeningmethoden eingegangen wird. Bei der Suche im Inhaltsverzeichnis findet man zur Prävention nur den Verweis zur geriatrischen Prävention, der einen siebenzeiligen Absatz aufweist und keinerlei grundlegende Erkenntnisse zur Prävention enthält [28]. Bei der weiterführenden Suche mit dem Schlagwort Screening wird man auf eine Seite der Psychosomatik weitergeleitet, bei der es lediglich um eine kurze Ausführung der heute verfügbaren Screeningfragebögen geht [41]. Ansonsten wird in den einzelnen Kapiteln ein Vermerk über die Möglichkeit und die Art der Durchführung eines Screenings gemacht, jedoch

finden sich keine genauen Angaben zu den Zahlenwerten in Bezug auf die Aussagekraft der jeweiligen Methode. Ein Standardwerk der Inneren Medizin das diesem Thema ein ganzes Kapitel mit dem Titel: „Screening und Prävention von Krankheiten“ widmet, ist Harrisons Innere Medizin. In diesem Kapitel werden die gängigsten Präventionsprogramme aufgezählt und bewertet. Es wird sowohl auf die Grundlagen des Screenings sowie auf die Endpunkte, die benötigt werden, um den Nutzen eines Screeningprogramms zu beurteilen, eingegangen. Es wird die NNS (Number needed to screen) am Beispiel für das Mammakarzinom und das Kolonkarzinom berechnet. Ein weiterer positiver Aspekt stellt die Tabelle 4-2 auf Seite 31 dar, welche die durchschnittliche Steigerung der Lebenserwartung einer Population durch Screeningmaßnahmen illustriert [44] und im Folgenden abgebildet ist.

Tabelle 8: Durchschnittliche Steigerung der Lebenserwartung einer Population (Abgewandelte Version aus Harrisons Innere Medizin, 17. Auflage, S. 31, Tabelle 4-2 [44])

Die folgende Tabelle verdeutlicht die durchschnittliche Steigerung der Lebenserwartung einer Population durch bestimmte Screeningmaßnahmen.
Auf der linken Seite der Tabelle werden die durchgeführten Screeningmethoden aufgezeigt und auf der rechten Seite die dazugehörige Steigerung der Lebenserwartung in Tagen, Monaten oder Jahren, die durch das jeweilige Screening erreicht werden können, illustriert.
(PSA: Prostata-spezifisches Antigen. PAP-Abstrich: Test zur Früherkennung eines Gebärmutterhalskrebses nach George Papanicolaou)

Screening	**Geschätzter Anstieg der Lebenserwartung**
Mammographie-Screening Frauen zwischen 40 und 50 Jahren Frauen zwischen 50 und 70 Jahren	 0-5 Tage 1 Monat
PAP-Abstrich, 18-65 Jahre	2-3 Monate
Belastungstest bei einem 50-Jährigen asymptomatischen Mann	8 Tage
PSA-Test und digitale rektale Untersuchung bei einem Mann über 50 Jahren	bis zu 2 Wochen

Aufgabe des Rauchens bei einem 35-Jährigen Raucher	3-5 Jahre
Durchführung regelmäßiger sportlicher Aktivitäten bei einem 40-Jährigen Mann (3x30 min pro Woche)	9 Monate bis 2 Jahre

Resultierend zeigt sich ein mageres Ergebnis bezüglich der Information über die Bewertung von Präventions- und Screeningprogrammen in wichtigen Lehrbüchern der Inneren Medizin. In den genannten Lehrbüchern wird die Existenz von Screeningmaßnahmen zwar aufgezeigt, wohingegen auf die Zuverlässigkeit, Probleme und die Aussagekraft solcher Tests nicht eingegangen wird. Ein Buch, das in dieser Beziehung wenigstens versucht einen kurzen Überblick über die Nachteile, die Kosteneffektivität und die durchschnittliche Lebensverlängerung, die durch ein Screening erwartet werden kann, zu geben, ist das Werk Harrisons Innere Medizin vom Jahre 2008.

3.2 Entwicklung eines Instruments zur Analyse von Screeningstudien

3.2.1 Biasformen in einem Screeningprozess

Ein Screeningprozess lässt sich in die drei Hauptschritte: Screening, Diagnostik und Therapie unterteilen. Für jeden dieser Teilschritte in diesem Prozess lassen sich Bias ableiten, die in der jeweiligen Phase eines Screeningablaufs in Erscheinung treten können. Bias sind systematische Fehler und verzerren „(…) ein Versuchsergebnis in eine bestimmte Richtung und verleiten zu fehlerhaften Schlüssen.“ [96, S. 256].

Ein Testverfahren in einer Studie ist nur dann valide, wenn der Test auch das misst, wofür er ursprünglich konzipiert wurde. Die Validität lässt sich in eine externe und interne Validität aufsplitten [81]. Die externe Validität betrifft die „Verallgemeinerbarkeit“ und „Übertragbarkeit“ [104, S. 254] einer Studie [104]. Die interne Validität beinhaltet die Verlässlichkeit einer Studie und wird durch das Auftreten von Bias und dem Maß an Präzision bestimmt [81].

Aufgrund der Analyse der Screeningstudie, die sich mit dem Screeningverfahren Low-Dose CT für die Früherkennung von Lungenkarzinom beschäftigte und infolge der Durchführung zahlreicher Diskussionen mit Doktoranden, die ebenfalls Arbeiten über Screeningstudien, unter der Leitung von Herrn Professor Porzsolt verfassten (Glaukom- und PSA-Screening), konnte die Abbildung 7 entwickelt werden. Hierbei wurden die möglichen Probleme, die während eines Screeningprozesses auftreten, zu den nachfolgenden Bias zusammengefasst und den drei großen Teilschritten dieses Prozesses untergliedert. Die Definitionen und Erklärungen zu den einzelnen Bias werden im Folgenden (Kapitel 3.2.2) dokumentiert.

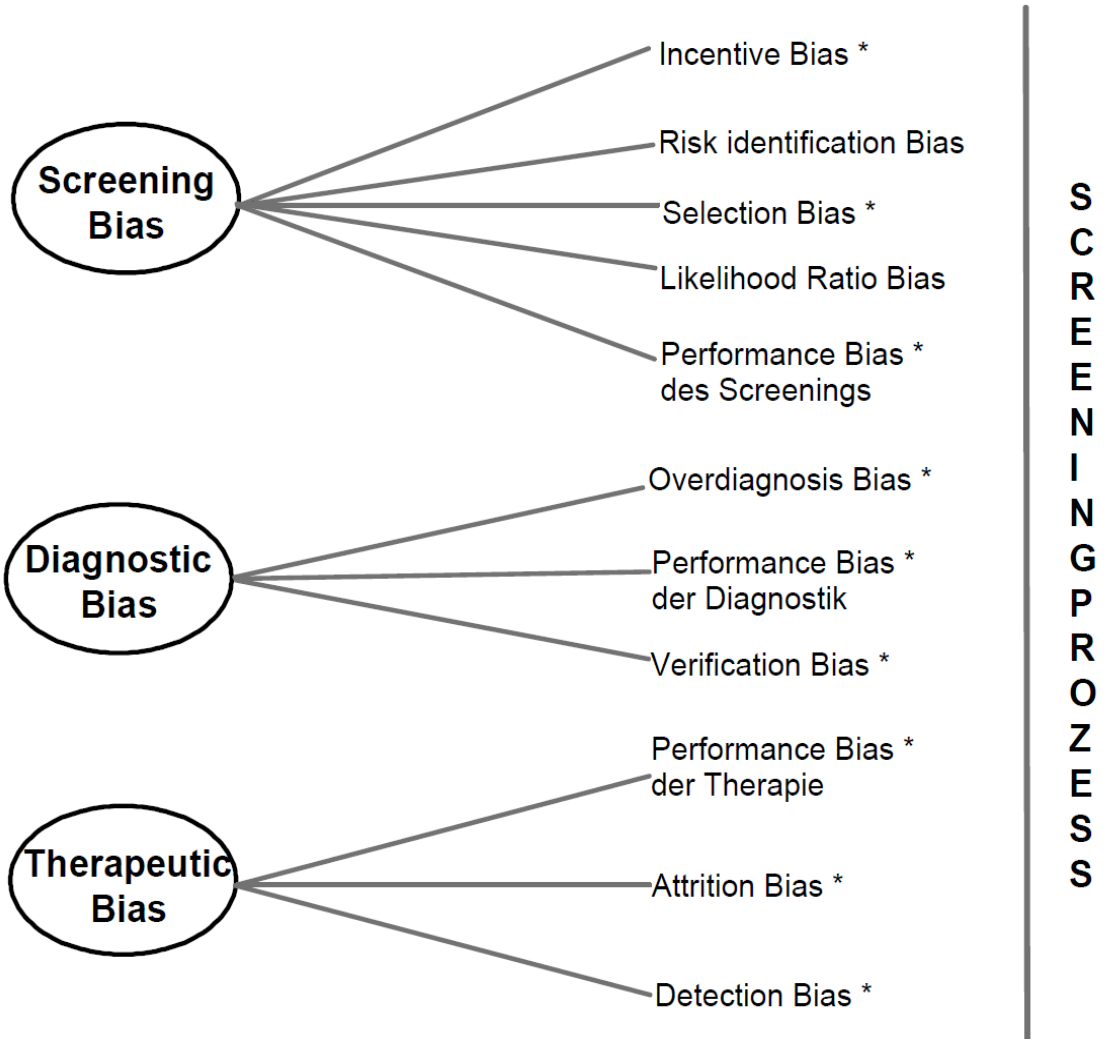

Abbildung 7: Bias in einem Screeningprozess

Die Arten von Bias, die im Verlauf eines Screeningprozesses auftreten können und denen für eine Studienanalyse Beachtung geschenkt werden sollten, werden den drei großen Teilschritten (Screening, Diagnostik und Therapie) zugeordnet.

(Bias: systematischer Fehler in einer Studie. Die Biasformen, die mit einem * markiert sind, existieren in der Literatur. Die Biasformen ohne * stellen eine Neuschöpfung dar.

Die genauen Definitionen und Zuordnungen der einzelnen Bias können den Ausführungen in Kapitel 3.2.2 entnommen werden.)

3.2.2 Definitionen und Erläuterungen der Biasformen in einem Screeningprozess

3.2.2.1 Screening Bias

Incentive Bias

Der Incentive Bias, der sich zu einer Untergruppe der Sampling Bias kategorisieren lässt, ist eine in den literarischen Werken weit verbreitete Form der Bias [42]. Als Sampling Bias wird eine Über- oder Unterrepräsentation von bestimmten Menschen mit demselben Merkmal (z.B. ethnische Rasse) in einer Studienpopulation bezeichnet, welches ein späteres Studienergebnis verfälschen und eine Generalisierung auf die Bevölkerung verwehren kann [10]. Nachdem sich einige Personengruppen durch Versprechungen eher für die Teilnahme an einer Studie begeistern lassen als andere, wird durch medienwirksame Werbungen versucht, Personen, die ein bestimmtes Risikoprofil tragen, anzusprechen. Dies geschieht zum Beispiel durch eine finanzielle Entlohnung oder durch weitere diagnostische Maßnahmen, die eine Vergleichspopulation in der Allgemeinbevölkerung nicht erhalten würde, auch wenn zu diesem Zeitpunkt die Überlegenheit der neuen Maßnahme nicht gesichert wäre [11]. Aufgrund der beschriebenen Schwierigkeiten besteht die Möglichkeit, dass der Incentive Bias dazu beitragen könnte, dass die Studienpopulation nicht den Durchschnitt der Allgemeinbevölkerung widerspiegelt, sondern aus einer selektierten Teilnehmergruppe besteht.

Nachdem einer wichtigen Untergruppe des Incentive Bias, dem Volunteer Bias, bei Studienanalysen besondere Aufmerksamkeit geschenkt werden muss, wird dieser von mir gesondert aufgeführt. Der Volunteer Bias oder auch Self-selection Bias entsteht durch den sogenannten „healthy volunteer effect“. Eine Studie von 2004, die sich mit dem Prostatakarzinom-Screening beschäftigte, konnte nachweisen, dass sich Studienteilnehmer, die sich freiwillig für eine Studie meldeten, nicht dem Durchschnitt der Zielpopulation entsprachen [57]. Dieser Personenkreis von Freiwilligen verfügte über ein besseres Gesundheitsbewusstsein, wie auch über einen höheren Bildungsstandard und war im Vergleich zu der Zielpopulation, für die die Studie kreiert wurde, gesünder [11][57].

Diese Effekte des Incentive Bias haben Auswirkungen auf die externe Validität einer Studie, da die Übertragbarkeit und Generalisierbarkeit auf die Zielbevölkerung erheblich eingeschränkt werden kann.

Risk identification Bias

Die Entstehung eines Risk identification Bias kann durch einen Confounder (Störfaktor) erfolgen. Ein Confounder ist eine mögliche dritte Variable, die fälschlicherweise eine Exposition mit einer Erkrankung in einen direkten Zusammenhang bringt [46][97].

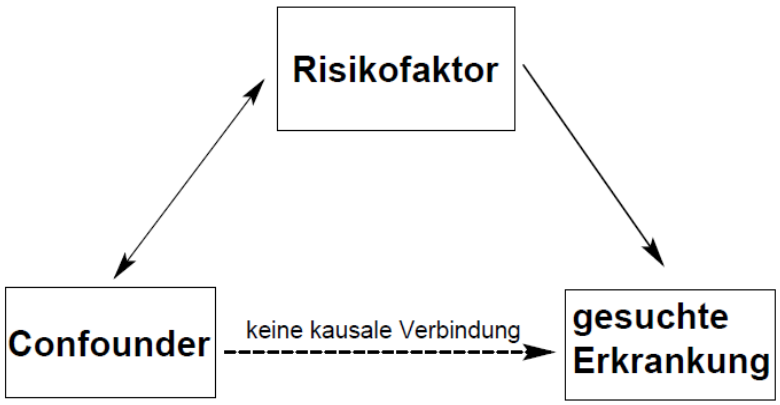

Abbildung 8: Risikofaktor und Confounder

(Abblidung nach „Concepts of Epidemiology“,Bhopal RS, 2002 [11])

Die obige Darstellung veranschaulicht die Problematik, dass ein Confounder fälschlicherweise für einen Risikofaktor einer Erkrankung gehalten werden kann, obschon er in keinem direkten Zusammenhang zur gesuchten Erkrankung steht.

(Zwischen dem Risikofaktor und der gesuchten Erkrankung besteht ein direkter kausaler Zusammenhang. Zwischen dem Confounder und der gesuchten Erkrankung besteht lediglich ein indirekter, aber kein kausaler Zusammenhang. Der Confounder und der Risikofaktor stehen in einem direkten Zusammenhang. Confounder: Störfaktor)

Wie in der oben beschriebenen Grafik dargestellt, kann nur eine korrekt gewählte Hochrisikogruppe, die einen Risikofaktor trägt, der direkt die gesuchte Erkrankung begünstigt, ein erfolgreiches Screening ermöglichen. Der Name Risk identification Bias wurde bisher in der Literatur nicht verwendet, jedoch wurde die Problematik erkannt, dass ein Confounder für einen Risikofaktor gehalten und so

Studienergebnisse verfälscht werden können [11]. Diese Biasform betrifft sowohl die externe als auch die interne Validität, da durch die Wahl eines inkorrekten Risikofaktors einerseits die Verallgemeinerbarkeit eingeschränkt (externe Validität) und andererseits das Studienergebnis vom wahren Wert abweichen (interne Validität) kann.

Selection Bias

Selection Bias entstehen durch Fehler in der Randomisation der Studienteilnehmer zu der Experimental- und Kontrollgruppe. Es treten „Unterschiede in der Ausgangssituation der Teilnehmergruppen (…)“ auf, die „(...) hinsichtlich der Studie – relevante Eigenschaften der Teilnehmer (...)“ darstellen [14]. Folglich sind die Kontrollgruppe und die Experimentalgruppe nicht mehr vergleichbar und Rückschlüsse auf Testergebnisse nicht mehr möglich.

Ein Selection Bias kann beispielsweise auftreten, wenn die Studienteilnehmer über den möglichen Nutzen und die Vorteile der zu untersuchenden Methode Bescheid wissen. Die Probanden nehmen an der Studie teil, um vom erwarteten Nutzen der neuartigen Methode zu profitieren. Nun erfolgt bei einer RCT (Randomised Controlled Trial) eine Randomisation in eine Experimental- und in eine Kontrollgruppe, wobei lediglich in der Experimentalgruppe das neue Verfahren getestet wird und die Kontrollgruppe die herkömmliche Untersuchung bekommt. Dies beinhaltet jedoch zum einen einen Placeboeffekt bei nicht verblindeten Mitgliedern der Experimentalgruppe, wie es auch das bekannte Experiment zum Placeboeffekt von Rebecca Waber und Kollegen 2008 zeigen konnte [95] und zum anderen sorgt es dafür, dass Studienteilnehmer in der Kontrollgruppe sich eher aus der Studie wieder ausschließen lassen oder auf privatem Wege Möglichkeiten suchen, dieselben diagnostischen Maßnahmen zu erhalten wie Teilnehmer aus der Experimentalgruppe. Diese Hoffnung führt dazu, dass sich die Mitglieder in der Experimentalgruppe wohler und besser aufgehoben fühlen als jene, die den Versprechungen glauben und sich nun in der Gruppe wiederfinden, die die beworbene effektivere Screeningmaßnahme nicht erhält. Den Placeboeffekt in dem experimentellen Arm und die Demotivation in dem kontrollierten Arm müssen bei werbenden Kampagnen für die Teilnahme an einer Studie mit in die nachfolgende reflektierende Diskussion miteinbezogen werden.

Diese Biasform steht dem Incentive Bias sehr nahe, da auch bei diesem Bias Anreize für die Teilnahme eine Rolle spielen. Der Unterschied besteht darin, dass hier das Hauptaugenmerk auf die Kontrollgruppe gelegt wird, denn diese kann versuchen das Studienprotokoll zu umgehen (Selection Bias). Beim Incentive Bias ähneln sich beide Gruppen in der Studie, indessen sie jedoch den Durchschnitt der Zielpopulation nicht widerspiegeln (Sampling Bias).
Der Selection Bias zeigt Auswirkungen auf die interne Validität einer Studie, da das Studienergebnis nicht mehr nur von der Güte des Tests abhängt, sondern auch von nicht kontrollierbaren Variablen (Eigenschaften der Teilnehmer, die möglicherweise die gesuchte Erkrankung begünstigen können) innerhalb der zwei Studienarme (Experimental- und Kontrollgruppe).

Likelihood Ratio Bias

Der Likelihood Ratio Bias kann auch als Sensitivitäts-Spezifitätsbias bezeichnet werden, wohingegen diese Biasform in der Literatur bisher nicht erwähnt wurde. Die Likelihood Ratio misst die Güte eines Tests und in diesem Fall die Güte des Screeningtests. Ein Likelihood Ratio Bias entsteht, wenn der Screeningtest unbrauchbar ist, das heißt wenn der Likelihood Ratio Quotient einen Wert nahe eins annimmt.
Der Likelihood Ratio Quotient (LHR) lässt sich wie folgt berechnen [98]:

positiver LHR (PLHR) = Sensitivität / (1-Spezifität)

negativer LHR (NLHR) = (1-Sensitivität) / Spezifität

Diese systematische Verzerrung tritt in Erscheinung, wenn die Screeningergebnisse nicht dichotom einem positiven oder negativen Befund zugeteilt werden können. Vor allem wenn die Screeningtests mit bildgebenden Verfahren durchgeführt werden, existiert kein Zahlenwert ab dem ein Befund als positiv bezeichnet werden kann. Bei vielen Screeningmaßnahmen, die mithilfe eines Ultraschalls, Computertomographien oder anderen bildgebenden Verfahren durchgeführt werden, obliegt es meist den untersuchenden Ärzten einen Befund als positiv oder negativ zu deklarieren. Bei Befunden, die nicht dichotom eingeteilt

werden können, sollten keine Maßnahmen erzwungen werden dies trotzdem zu ermöglichen, denn somit steigt das Risiko des Likelihood Ratio Bias an. Damit das neue Screeningverfahren für die Allgemeinheit anwendbar ist, bedarf es eindeutigen Empfehlungen, an denen sich die Ärzte bei der Bewertung des Bildmaterials orientieren können. Ansonsten wird es für spätere Institutionen schwierig werden, dieses Screening erfolgreich durchzuführen, wenn keine klaren Vorgaben existent sind, wie Befunde zu klassifizieren sind (interne Validität). Dadurch entstehende negative Folgen, können sich in ansteigenden Raten von falsch positiven und negativen Befunden widerspiegeln, die eine weitere kostspielige und mögliche komplikationsreiche Diagnostik für die betroffenen Patienten bedeutet.

Ein weiterer wichtiger Punkt, um die Güte eines Tests einschätzen zu können, ist der Ausbildungsstand der Ärzte in der Studie. Eine Schwierigkeit besteht dann, wenn die Ärzte in der Allgemeinbevölkerung einen schlechteren Ausbildungsstandard aufweisen, als die vergleichbaren Ärzte in der Studie. Diese Thematik schränkt die Übertragbarkeit der Studienergebnisse auf andere Kliniken ein und darf nicht übersehen werden (externe Validität).

Zusätzlich sollte darauf geachtet werden, ob die Ärzte, die das Screening durchführen, Einsicht in die Patientenakten und damit Informationen zum klinischen Gesundheitszustand ihrer Patienten erhalten haben. Durch klinische Vorinformationen zu den Patienten kann die richtige Trefferquote für Screeningbefunde künstlich erhöht werden. Dies tritt vor allem dann auf, wenn der Test keinen definierten Wert liefert, ab welchem ein Befund als positiv oder negativ zu klassifizieren ist. Es bleibt daher offen, inwieweit das Screening allein eine Reduktion der Mortalität erreicht und in welchem Ausmaß das klinische Wissen über den Patienten die eigentliche Diagnose erbringt. In dieser Situation ist es nicht mehr möglich den Test alleine zu beurteilen, sondern es wird auch der durchführende Arzt bewertet.

In diesem Kontext stellt die Verblindung des beteiligten medizinischen Personals einen zentralen Punkt dar. Schon allein die Tatsache, dass sie die Zuordnung der Patienten zu den Gruppen kennen, lässt sie möglicherweise eine andere Verhaltensweise den Personen in der Kontrollgruppe zugrunde legen als gegenüber den Personen in der Experimentalgruppe (Interviewer Bias) [79].

Demgegenüber bedeutet keine Verblindung der Ärzte auch den möglichen Zugriff auf die medizinischen Daten der Patienten und damit eine Erhöhung der Trefferquote für richtige Befunde (Beeinträchtigung der internen Validität).
Diese Biasform (Likelihood Ratio Bias) beeinflusst die interne und externe Validität der Screeningstudie.

Performance Bias des Screenings

„Ein Performance Bias liegt vor, wenn die untersuchten Gruppen unterschiedlichen Maßnahmen (mit Ausnahme der geprüften Intervention) ausgesetzt sind, welche das Ergebnis beeinflussen können.“ [67, S.33]. Existieren innerhalb einer Studie Unterschiede in den Rahmenbedingungen, die sowohl zwischen den teilnehmenden Studienzentren als auch zwischen Kontroll- und Experimentalgruppe auftreten können, kann der oben benannte Fehler in Erscheinung treten. Diese Abweichungen führen zur Inkongruenz innerhalb der Studienpopulation und erschweren eine genaue Beurteilung der Studienergebnisse. Dieser Bias kann umgangen werden, wenn die randomisierte Studie doppelt blind (sowohl Patienten als auch Ärzte sind verblindet) durchgeführt wird [67] und sich die beteiligten Institutionen bei der Durchführung exakt ans Studienprotokoll halten. Der Performance Bias kann bei der Durchführung von Tests (Screening- und Diagnostiktests) und bei Therapiemaßnahmen auftreten und beeinflusst primär die interne und sekundär die externe Validität einer Studie.

3.2.2.2. Diagnostic Bias

Overdiagnosis Bias

Im Rahmen diagnostischer Maßnahmen kann es zum Auftreten eines Overdiagnosis Bias kommen, der die interne Validität beeinträchtigt.
In literarischen Werken wird der Begriff „Überdiagnose“ sehr häufig und aktuell vor allem im Zusammenhang mit dem Mammografie- und PSA-Screening verwendet [69][47][21].
Durch die Teilnahme an einer Screeningstudie erhoffen sich die Teilnehmer ein umfassendes medizinisches Engagement, wie bereits bei dem Punkt Incentive

Bias besprochen. Diese Studie induziert bei einigen Teilnehmern eine gefühlte Sicherheit und die Bereitschaft, jegliche diagnostische Maßnahme im Falle eines positiven Befundes des Screeningtests in Kauf zu nehmen, wenngleich der Screeningtest noch den Nachweis über seine Wirksamkeit schuldig bleibt. Bei einigen dieser Patienten wird ein Karzinom nachgewiesen und daraufhin behandelt, obwohl dieses Karzinom sie in ihrem Leben zu keiner Zeit beeinträchtigt hätte [69]. Diese entdeckten Karzinome, die für die Betroffenen nie eine Relevanz in ihrem Leben gespielt hätten, werden zugunsten des Screeningprogramms gewertet, da diese Karzinome langsam und weniger invasiv wachsen und damit eine gute 5-Jahres-Überlebensrate aufweisen. Dies führt zu einer Überdiagnostik und zu unnötigen personellen und finanziellen Belastungen für unser Gesundheitswesen und zu möglichen Komplikationen bis hin zu Todesfällen für die Betroffenen [43].

Darüber hinaus besteht beim Screening die Gefahr, dass einige der Teilnehmer sich zusätzlichen diagnostischen Maßnahmen außerhalb der Studie unterziehen. Dieses Problem tritt in Erscheinung, wenn den Patienten ihr Screeningbefund mitgeteilt wird. Handelt es sich um ein Indifferenzstadium (Zwischenstadium), das zunächst keiner weiteren Diagnostik bedarf, sondern lediglich in einer der folgenden Screeningrunden erneut bewertet wird, so werden einige dieser Personen weitere diagnostische Maßnahmen außerhalb der Studie durchführen lassen, um nicht länger in Ungewissheit leben zu müssen. Auch diese Situation führt zu einem Ressourcenverbrauch im Gesundheitswesen und kann nur durch eine Verblindung der Patienten bezüglich ihres Testergebnisses umgangen werden.

Performance Bias der Diagnostik

Der Performance Bias der Diagnostik wurde bereits bei den Screening Bias besprochen (siehe dazu Performance Bias des Screenings). In diesem Fall wird der Performance Bias nicht auf den Screeningtest, sondern auf den diagnostischen Test bezogen.

Verification Bias

Der Verification Bias wird im Buch „Chirurgische Forschung“ von M. Krukemeyer und H.-U. Spiegel aufgeführt [79]. Als Verification Bias wird eine differenzielle Verifikation bezeichnet, die vorherrscht, wenn das Ergebnis des Screeningtests nicht bei jedem Studienteilnehmer mit einem anschließenden diagnostischen Test überprüft wird. Das Resultat einer Screeningmaßnahme, insbesondere wenn es sich um ein Karzinom-Screening handelt, kann oft nur mittels einer Biopsie bestätigt werden. Ausschließlich nach Erhalt eines positiven Screeningresultats wird bei den Patienten eine Gewebebiopsie vorgenommen. Folglich können nur die falsch positiven, nicht aber die falsch negativen Resultate ermittelt werden [40]. Aus einer ausschließlichen Nachverfolgung von positiven Screeningergebnissen resultiert, dass die Sensitivität des Screeningtests falsch niedrig und die Spezifität falsch hoch angenommen wird. Der Verification Bias weist eine Beeinflussung der internen Validität auf.

3.2.2.3 Therapeutic Bias

Performance Bias der Therapie

Die Definition und Erläuterung dieses Bias wurde bei Performance Bias des Screenings (Kapitel 3.2.2.1) angegeben. Bei diesem Punkt wird der Performance Bias nicht auf den Screeningtest, sondern auf die durchgeführte Therapie bezogen.

Attrition Bias

Ein Attrition Bias ähnelt einem Selection Bias, der durch das Ausscheiden von Studienteilnehmern, die ein Ungleichgewicht innerhalb der Studiengruppen bewirken, verursacht wird [67]. Hewitt und Hahn et al. untersuchten in ihrem Review Studien, die in den vier großen medizinischen Journals (BMJ, JAMA, Lancet, NEJM) im Zeitraum von Januar bis Dezember 2002 erschienen waren. Gleichzeitig prüften sie, ob alle Studienteilnehmer nachbeobachtet wurden und eine genaue Dokumentation dieser erfolgte. Trotz der Aufnahme des CONSORT Statements (in Kapitel 2.2 beschrieben) in einigen Studien, mangelte es bei 54% von 132 Studien in einer adäquaten Nachverfolgung und lückenlosen

Dokumentierung der Teilnehmer, die die Studie verlassen hatten [33][20]. Ein etabliertes Verfahren, um das Ausscheiden von Studienteilnehmern korrekt zu analysieren, ist die Intention to screen/treat Analyse. Bei dieser Methode werden die ausgeschiedenen Patienten in der Gruppe ausgewertet, zu der sie ursprünglich randomisiert und zugeordnet wurden [35]. Bei einer Untersuchung von Veröffentlichungen aus dem Jahre 1997, publiziert in denselben Journals, die Hewitt und Hahn benutzten, wird die Intention to treat Analyse häufig (48%) erwähnt, jedoch unzulänglich gebraucht [35].

Die Personen, die eine Studie verlassen, sollten nach ihrem Erfolg in der Studie bewertet werden, um zu vermeiden, dass nicht diejenigen aus der Studie ausscheiden, bei denen die Therapie oder das Screening versagte oder mögliche Nebenwirkungen in Erscheinung getreten sind. Des Weiteren muss auf das Gleichgewicht der Personeneigenschaften (Ethnische Herkunft, Geschlecht, Alter und andere Faktoren) der Studienabbrecher aus der Kontroll- und der Experimentalgruppe geachtet werden [67]. Der Attrition Bias zeigt Auswirkungen auf die interne Validität [20].

Detection Bias

Der Detection Bias wird in der Literatur häufig verwendet und umfasst eine Reihe von Punkten, denen Beachtung geschenkt werden sollte [67]. Zum einen sollte der Endpunkt einer Studie eindeutig definiert sein und zum anderen sollte das Verfahren um diesen Endpunkt festzulegen, kritisch hinterfragt werden. Wird als Endpunkt die disease specific mortality (die Erkrankung verursacht den Tod des Patienten) gewählt, gestaltet es sich diffizil, die genaue Todesursache festzulegen, da oft nicht eindeutig geklärt werden kann, ob der Patient an den Folgen seiner Karzinomerkrankung oder beispielsweise an einer Erkrankung des Herz-Kreislaufsystems gestorben ist. Auf diese Problematik verwies auch ein Artikel aus dem Journal of medical screening aus dem Jahre 2010 [38], in dem festgestellt wurde, dass die disease specific mortality Rate wohl den Effekt des Prostatakarzinom-Screenings unterschätzt hatte.

Des Weiteren bedarf es einer adäquaten Dauer der Nachverfolgungszeit der Studienteilnehmer, um den tatsächlichen Benefit einer Screeningmaßnahme auch sichern zu können.

Zusätzlich sollte darauf geachtet werden, inwieweit die Komplikationsraten therapeutischer Interventionen, sowie die peri- und postoperative Sterblichkeit den aktuellen Zahlen entsprechen oder ob Therapieergebnisse zu Gunsten eines besseren Studienoutcomes geschönt wurden. Diese Reihe an Anmerkungen beeinflussen das Ergebnis einer Studie (Auswirkungen auf die interne Validität) und führen zu einer minimierten Reproduzierbarkeit der Studienergebnisse im klinischen Alltag (Auswirkungen auf die externe Validität).

3.2.3 Einteilung der Bias in direkte und indirekte Bias

Die Einteilung einer Screeningstudie in die drei großen Bias: Screening-, diagnostic- und therapeutic Bias umfasst nicht alle wichtigen Details einer Studie. Bei der Analyse der Studie aus dem NEJM und durch die Diskussionen mit anderen Doktoranden, die den Nutzen des Glaukom- und des Prostatakarzinom-Screenings bewerteten und durch die Anwendung des erweiterten USP-Fragebogens (siehe Anhang, Tabelle 7), der bereits im Methodenteil vorgestellt wurde, konnten weitere wichtige Komponenten herausgefunden werden, denen man bei der Evaluation von Screeningstudien größere Aufmerksamkeit zukommen lassen sollte. Diese Items wurden schließlich zu den indirekten Bias zusammengefasst. Diejenigen Bias, die im Verlauf eines Screeningprozesses auftreten, wurden unter dem Begriff „direkte Bias“ zusammengefasst. Die anderen Punkte, die zusätzlich für die Analyse von Screeningstudien relevant sind, wurden unter dem Begriff „indirekte Bias“ subsumiert. Dabei bedeutet indirekte Bias nicht, dass diese Themen weniger Auswirkungen auf die Validität und klinische Relevanz einer Studie haben, sondern lediglich, dass sie nicht direkt aus dem Screeningablauf resultieren.

Tabelle 9: Direkte und indirekte Bias

In der folgenden Tabelle werden die Bias, die in einer Screeningstudie auftreten können und bedeutungsvoll für die Analyse jener ist, in direkte und indirekte Bias aufgegliedert. Die direkten Bias treten im Verlauf eines Screeningprozesses auf. Die indirekten Bias stellen zusätzliche Fehlerquellen dar, die nicht direkt im Verlauf eines Screeningprozesses auftreten, aber ein Studienergebnis verzerrt darstellen können.

Auf der linken Seite der Abbildung werden die Bias, die in Kapitel 3.2.2 aufgeführt und erläutert wurden, zu den direkten Bias zusammengefasst. Auf der rechten Seite der Abbildung finden sich zusätzliche Items, die dem erweiterten USP-Fragebogen (siehe Anhang, Tabelle 7) entnommen wurden und zu den indirekten Bias zusammengefasst wurden.

(Bias: systematischer Fehler. USP: Usability of scientific publication)

DIREKTE BIAS	INDIREKTE BIAS
Screening Bias	> Wahl des idealen Studiendesigns
	> Geheime Allokation zu den Studiengruppen
Diagnostic Bias	> Verblindung der Ärzte und der Patienten
	> Akzeptabler Vergleich (falls vorhanden) zu einem Indextest
	> Verteilung der Krankheitsstadien bei Diagnosestellung
	> Akzeptable Nebenwirkungen des Screeningtests für die Patienten
	> Verfügbarkeit des neuen Screeningtests
Therapeutic Bias	> Existenz effektiver Therapien
	> Ausschluss von Interessenskonflikten
	> Angemessene Statistik

Im folgenden Kapitel (3.3) wird diese tabellarische Darstellung der direkten und indirekten Bias auf die Beispielstudie „Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening“ aus dem New England Journal of Medicine aus dem Jahre 2011 angewendet.

3.2.4 Fragensammlung für die Evaluation von Screeningstudien

Für die Evaluierung von Screeningstudien wurde eine Sammlung von Fragen entwickelt, die aus den direkten und indirekten Bias besteht. Die direkten und indirekten Bias wurden zu kurzen und präzisen Fragen formuliert, wodurch das Analyseinstrument einfacher in der Handhabung wird und somit die Leser einer Studie die detailliert gestellten Fragen beantworten können, um eine Vorstellung darüber zu erhalten, ob die Screeningstudie über eine ausreichend hohe interne und externe Validität und über eine klinische Relevanz verfügt.

Tabelle 10: Fragensammlung für die Evaluierung von Screeningstudien

Dieser Fragebogen wurde für die Analyse von Screeningstudien konzipiert, um Lesern einer Studie ein Hilfsmittel zu geben, Screeningmaßnahmen bewerten zu können.

In der linken Spalte erfolgt die Zuordnung der einzelnen Fragen zu den Bias. Die zweite Spalte lässt die Zuordnung der Fragen zu ihrer Nummerierung erkennen. Diese Nummerierung wird auch in Kapitel 3.3 beibehalten. In der dritten Spalte werden die Fragen zu den Bias formuliert und in den beiden letzten Spalten können die Antwortmöglichkeiten entweder zu „Ja“ (zutreffend) oder „Nein“ (nicht zutreffend) gegeben werden.

(Bias: systematischer Fehler. Direkte Bias: treten im Verlauf eines Screeningprozesses auf. Indirekte Bias: zusätzliche Fehlerquellen, die nicht direkt im Verlauf eines Screeningprozesses auftreten, aber ein Studienergebnis verzerrt darstellen können. Confounding: verzerrte Darstellung von Studienergebnissen aufgrund der Existenz eines Störfaktors. Goldstandard:aktuell bester verfügbarer Test zum Nachweis der Erkrankung. Intention to screen: Auswertung der Teilnehmer in der Gruppe, zu der sie ursprünglich randomisiert wurden. Length time bias: Langsam wachsende Tumore werden in einem Screeningverfahren eher erkannt, da sie für längere Zeit ein symptomfreies Intervall aufweisen als schnell wachsende Tumore.)

Bias	**Nr.**	**Fragen zur Evaluation von Screeningstudien**	**Antwortmöglich–keiten** (mit genauer Begründung und Textangabe)	
Direkte Bias			Ja, weshalb?	Nein, weshalb?
Incentive Bias	1	Wurde die Teilnehmerrekrutierung durch bestimmte Anreize getriggert?		
Risk identification Bias	2	Erfolgte eine geeignete Wahl der Risikofaktoren, die die gesuchte Erkrankung auch direkt begünstigen können, um die Möglichkeit eines Confoundings auszuschließen?		
Selection Bias	3.1	Bestehen Unterschiede - in für die Studie -relevanten Personeneigenschaften bei den zu vergleichenden Gruppen?		
	3.2	Erfolgte eine Analyse zur Präferenz des Screeningverfahrens bei den Teilnehmern?		
Likelihood	4.1	Ist die Güte des zu untersuchenden Screeningtests ausreichend hoch?		

Ratio Bias	4.2	Existieren eindeutige Empfehlungen ab wann ein Screeningbefund als positiv oder negativ zu bewerten ist?		
	4.3	Unterscheiden sich die teilnehmenden Ärzte in der Studie durch einen höheren Ausbildungsstand von denen außerhalb der Studie?		
	4.4	Hatten die behandelten Ärzte die Möglichkeit, sich klinische Informationen der Patienten anzueignen?		
Perfor-mance Bias	5	Existieren Unterschiede in den Rahmenbedingungen, die das Screening, die Diagnostik oder die Therapie betreffen, sowohl zwischen den zu vergleichenden Gruppen als auch zwischen den teilnehmenden Studienzentren?		
Overdia-gnosis Bias	6	Sind die Probanden bezüglich ihres Testergebnisses verblindet? Falls nein, besteht die Möglichkeit der zusätzlichen Inanspruchnahme von Diagnostik außerhalb der Studie? Wie ist das Problem der möglichen Überdiagnose von der gesuchten Erkrankung einzuschätzen?		
Verifica-tion Bias	7	Wird bei allen Patienten das Screeningergebnis anhand eines Goldstandards verifiziert?		
Attrition Bias	8.1	Erfolgte eine „Intention to screen“ Analyse?		
	8.2	Sind die Studiengruppen auch noch nach dem Ausscheiden von Studienteilnehmern bezüglich ihrer Personeneigenschaften ähnlich?		
Detection Bias	9.1	Sind die Endpunkte der Studie genau definiert? Sind die Dimensionen definiert in welchem		

		die Ziele gemessen werden sollen?		
	9.2	Ist die Nachbeobachtungszeit ausreichend lange, um adäquate Endpunkte zu messen?		
	9.3	Entsprechen die therapeutischen Komplikationsraten innerhalb der Studie den aktuellen Zahlen außerhalb der Studie?		
Indirekte Bias	10	Wurde das ideale Studiendesign gewählt?		
	11	Erfolgte die Allokation zu den Studiengruppen geheim?		
	12	Waren die Ärzte und die Patienten verblindet?		
	13	Ist bei Studien im Parallelgruppen-Design der zu vergleichende Indextest akzeptabel?		
	14	Können aus der Verteilung der Schweregrade der Erkrankung bei Diagnosestellung Rückschlüsse auf einen Length time bias gezogen werden? Was fällt zusätzlich noch auf?		
	15	Sind die Nebenwirkungen des neuen Screeningtests für die Patienten akzeptabel?		
	16	Ist der Screeningtest in der zukünftig zu untersuchenden Gesellschaft ausreichend verfügbar?		
	17	Bestehen effektive Therapiemöglichkeiten für die gescreente Erkrankung?		
	18	Bestehen Interessenskonflikte?		
	19	Ist die Statistik angemessen?		

3.3 Anwendung der vorgeschlagenen Evaluationsmethode auf die Screeningstudie „Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening“

Die entwickelte Fragensammlung für Screeningstudien wird in diesem Kapitel auf eine Studie aus dem New England Journal of Medicine mit dem Titel „Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening“ aus dem Jahre 2011 [1] angewendet. Bei dieser Anwendung kann gezeigt werden, dass mithilfe dieser Fragensammlung eine Screeningstudie bezüglich ihrer Validität, Reliabilität und klinischen Relevanz eingeschätzt werden kann, wobei zugleich ersichtlich wird, dass der veröffentlichte Studienbericht allein oft nicht ausreicht, um die gestellten Fragen adäquat beantworten zu können.
Der folgenden Auflistung an Fragen liegt die Fragensammlung und die Erläuterungen zu den angewendeten Bias aus dem Kapitel 3.2 zugrunde.

Direkte Bias

(Frage 1)
Wurde die Teilnehmerrekrutierung durch bestimmte Anreize getriggert?
(Incentive Bias)

Ein Incentive Bias tritt auf, wenn durch Anreize bestimmte Personen mit bestimmten Merkmalen sich eher in die Studienpopulation aufnehmen lassen und somit Unterschiede zur Zielpopulation bestehen, für die der zu untersuchende Test konzipiert wurde.
Die Rekrutierung der Hochrisikogruppen für die Studie erfolgte auf freiwilliger Basis, welche durch die NCI Office of Communication geleitet wurde. Es wurden unter anderem folgende Strategien verfolgt [3, S.16]:

- Ausstrahlung einer Werbesendung im Satellitenfernsehen
- Einrichtung einer Telekonferenz, bei der die Medienunternehmen Fragen und Antworten an die Hauptverantwortlichen des NLST stellen konnten

- Schaltung regionaler Anzeigen
- Informierung verschiedener Ärzte, Kliniken und bestehende Gesundheitsprogramme über die Ziele der Studie, mit dem Gesuch Mithilfe bei der Rekrutierung von passenden Patienten zu leisten
- Gezielte Versendung von Mails in Stadtteile, die einen hohen Prozentsatz einer bestimmten ethnischen Gruppe enthielten
- Benutzung bereits vorhandener Hochrisikogruppen von anderen Studien, wie zum Beispiel die LHS (Lung Health Study) und dem NETT (National Emphysema Treatment Trial)

Der Incentive Bias, wie in Kapitel 3.2 beschrieben, betrifft die externe Validität einer Studie, da die rekrutierten Personen erst nach der Einwilligung zur Teilnahme an der Studie zu den beiden Studienarmen randomisiert wurden.
Die vorgenommene Rekrutierung birgt zudem das erhebliche Risiko des sogenannten „healthy volunteer“ Effekts. Gesundheitsbewusste Bürger werden eher an einer Studie teilnehmen, um möglichst früh eine drohende Erkrankung zu erkennen. Der Anreiz spiegelt sich für diese Personen in der Screeningmaßnahme wider, denn vor der Durchführung der Studie wurde kein Lungenkarzinom-Screening implementiert. Dieses Klientel der „healthy volunteers“ unterscheidet sich in mehreren Kriterien von der Allgemeinheit. Einerseits weisen Freiwillige an solchen Studien einen sehr viel bewussteren Lebensstil auf und sind gesünder und andererseits kann diese Personengruppe oft Frühzeichen einer Krankheit besser erkennen, da sie belesen und informiert ist [61].
Aus diesem Grund kann das ausgewählte Patientenklientel nicht die durchschnittliche Bevölkerung repräsentieren. Die Existenz eines Incentive Bias muss in dieser Studie berücksichtigt werden.

(Frage 2)

Erfolgte eine geeignete Wahl der Risikofaktoren, die die gesuchte Erkrankung auch direkt begünstigen können, um die Möglichkeit eines Confoundings auszuschließen?
(Risk identification Bias)

Der Risk identification Bias entsteht, wenn fälschlicherweise ein direkter Zusammenhang zwischen einem vermeintlichen Risikofaktor und einer Erkrankung angenommen wird und dadurch die Ergebnisse eines Screeningtests nicht verwertbar sind.

Es gab eine Reihe von Kriterien, die die Teilnehmer erfüllen mussten, um in der Studie teilnehmen zu können. Die Einschlusskriterien beinhalteten unter anderem:

1. Alter zwischen 55 und 74 Jahren
2. Eine Rauchervergangenheit von mindestens 30 pack-years (Pack-years ergeben sich aus der Multiplikation von gerauchten Zigaretten-schachteln pro Tag mit der Anzahl der Jahre, in denen geraucht wurde)
3. Ex-Raucher, die innerhalb der letzten 15 Jahre das Rauchen aufgehört hatten
4. Fähigkeit mit den Armen über dem Kopf liegen zu können
5. Eigenständige Unterzeichnung des Informed Consent

Außerdem wurden elf Ausschlusskriterien formuliert [4, S. 250]. Detailliertere Ein- und Ausschlusskriterien sind im Studienprotokoll dokumentiert [3, S. 4-18].

Die meisten Lungenkarzinome treten in einem Alter zwischen 50 und 80 Jahren auf [22], wobei die Inzidenz (Relative Anzahl der Neuerkrankungen in einer Bevölkerung in einem bestimmten Zeitraum) in Deutschland bei den 75- bis 80-Jährigen am höchsten ist [29]. Damit wurde das erste Einschlusskriterium nur teilweise korrekt gewählt, das besagte, dass nur Patienten in die Studie aufgenommen werden, deren Alter zwischen 55 und 74 Jahren läge.

In einem kürzlich erschienenen Review aus dem Deutschen Ärzteblatt [7] wird bestätigt, dass das Lungenkarzinom schätzungsweise zu 90 Prozent durch das Inhalieren von Zigarettenrauch verursacht wird. Das Rauchen steht folglich in einem direkten Zusammenhang mit dem Auftreten eines Karzinoms der Lunge. Auch ehemalige Raucher weisen noch nach Jahrzehnten ein höheres Risiko für eine maligne Entartung in der Lunge auf als Nichtraucher desselben Alters [60]. Die beiden Einschlusskriterien (2. und 3.) sind plausibel.

Da die Patienten in der Experimentalgruppe insgesamt dreimal eine CT-Untersuchung bekamen, mussten sie in der Lage sein, eine kurze Zeit mit den Armen über dem Kopf liegen zu können, da ansonsten eine Untersuchung dieser

Art nicht oder nur mit wesentlich schlechteren Ergebnissen vorgenommen hätte werden können (Einschlusskriterium vier).

Patienten mussten aufgrund ihres Persönlichkeits- und Selbstbestimmungsrechts über die zu durchführende Maßnahme, mögliche weiterführende Untersuchungen und die daraus entstehenden psychischen und physischen Folgen und Komplikationen aufgeklärt werden (Informed Consent, Einschlusskriterium fünf).

Das erste Einschlusskriterium hätte bei alleiniger Berücksichtigung der Inzidenz eines Lungenkarzinoms Patienten in einem Alter zwischen 55 und 80 Jahren einschließen müssen. Dies ist jedoch nicht der einzige Parameter, der beachtet werden muss. Eine ethische Berechtigung für eine Screeningmaßnahme mit kurativ intendierten Operationen für eine Karzinomerkrankung jenseits des 75. Lebensjahres wäre fraglich. Die anderen Einschlusskriterien für die Erlaubnis zur Teilnahme an der Studie wurden korrekt gewählt.

(Frage 3.1)

Bestehen Unterschiede in – für die Studie – relevanten Personeneigenschaften bei den zu vergleichenden Gruppen?
(Selection Bias)

Ein Selection Bias hat Auswirkungen auf die interne Validität einer Studie, da die zu vergleichenden Studienarme Unterschiede in Bezug auf Personenmerkmale aufweisen und somit die beiden Studiengruppen nicht vergleichbar sind.

Wie im Studienbericht aus dem NEJM in der Tabelle 1 auf Seite 399 zu erkennen ist [1], waren die Charakteristiken der Teilnehmer in den beiden Gruppen annähernd gleich. Dennoch fällt ein wichtiges Detail ins Auge, wenn man sich die Stadienverteilung der Lungenkarzinome, die zu den jeweiligen Screeningrunden und zu den Studienarmen aufgeschlüsselt sind, betrachtet [2, Tabelle 2, S. 38]. Die Screeningrunde T0 wurde direkt nach der Randomisation der Probanden in die beiden Studienarme durchgeführt. Zu diesem Zeitpunkt T0 befanden sich bereits 25% der Lungenkarzinome in der Röntgen-Gruppe im Stadium IV, während es in der CT-Gruppe nur 16% waren [2, Tabelle 2]. Tumorpatienten im Stadium IV weisen eine deutlich schlechtere Überlebenswahrscheinlichkeit und geringere Heilungschancen auf [29]. Dies lässt die Schlussfolgerung zu, dass bereits zu

Beginn der Studie Unterschiede zwischen den Gruppen bestanden (Selection Bias). Dies wiederum führt zu der Annahme, dass die Röntgen-Gruppe im Vergleich zur CT-Gruppe eine schlechtere Ausgangssituation besaß, da zum Zeitpunkt T0 bereits 56% mehr Patienten der Röntgen-Gruppe ein Stadium IV eines Lungenkarzinoms verglichen mit der CT-Gruppe aufwiesen. Diese Selektion ist weitgehend unabhängig von der durchgeführten Screeningmaßnahme.

(Frage 3.2)

Erfolgte eine Analyse zur Präferenz des Screeningverfahrens bei den Teilnehmern?
(Selection Bias)

Nicht verblindete Teilnehmer können, falls sie in den präferierten Studienarm randomisiert werden, einen Placeboeffekt aufweisen. Im Gegensatz dazu können die anderen Teilnehmer, die sich im unerwünschten Studienarm befinden, zusätzliche Diagnostik außerhalb der Studie in Anspruch nehmen, um dieselben Maßnahmen zu erhalten, die sie ursprünglich erhofft hatten. Damit unterscheiden sich die beiden Studiengruppen und ein Selection Bias tritt in Erscheinung, der die interne Validität beeinflusst.
In dem Studienbericht aus dem NEJM gab es keine Angaben dazu, welches Screeningverfahren die Teilnehmer bevorzugten [1]. Auch im dazugehörigen Studienprotokoll [3] ließen sich keine Andeutungen finden, dass die Teilnehmer nach diesem Detail gefragt wurden.

(Frage 4.1)

Ist die Güte des zu untersuchenden Screeningtests ausreichend hoch?
(Likelihood Ratio Bias)

Bei einem Screeningtest ist eine möglichst hohe Spezifität oder eine hohe positive Likelihood Ratio anzustreben, um diejenigen Patienten zu erkennen, die keiner weiterführenden Diagnostik bedürfen. Ein Test mit hoher Sensitivität birgt ein erhebliches Risiko für die Patienten, da die Rate an falsch positiven

Testergebnissen hoch sein wird und so viele der positiv eingestuften Teilnehmer einer unnötigen Diagnostik unterzogen werden.

Die Sensitivität und die Spezifität lassen sich durch eine Vierfeldertafel für jede Screeningrunde (T0, T1, T2) nur einzeln anfertigen, da ein Patient einen Umfang von mindestens einem (z.B. Diagnose Lungenkarzinom bei T0) bis höchstens drei Befunden (z.B. Patient leidet nicht an der gesuchten Erkrankung) aufweisen kann. Insgesamt werden für die CT- und Röntgenuntersuchungen sechs Vierfeldertafeln benötigt.

Aus dem Studienbericht lassen sich für die Screeningrunde T0 folgende Zahlenwerte für das CT und Röntgen finden:

Gesamtzahl der Gescreenten im CT: 26309

Gesamtzahl der Gescreenten im Röntgen: 26035. [1, S. 400, Tabelle 2]

Positive Screeningresultate im CT: 7191, von diesen Personen haben 270 (TP) ein Lungenkarzinom.

Positive Screeningresultate im Röntgen: 2387, von diesen haben 136 (TP) ein Lungenkarzinom. [1, S. 401, Tabelle 3].

Aus den genannten Zahlenwerten lassen sich die falsch Positiven (FP) berechnen. Das sind für die CT-Untersuchung 6921 (Differenz aus 7191 und 270) und für die Röntgenuntersuchung 2251 (Differenz aus 2387 und 136). Analog dazu lässt sich die Gesamtzahl an negativen Testergebnissen berechnen und zwar sind das für das CT 19118 (26309-7191) und für das Röntgen 23648 (26035-2387). Es fehlen in dem Studienbericht aus dem NEJM Zahlenangaben zu den aufgetretenen Lungenkarzinomen (TP+FN) nach der ersten Screeningrunde, jedoch lässt sich die Zahl an Lungenkarzinomen nach der Runde T0 aus der Tabelle 2 aus dem Supplementary Appendix des Studienberichts herausfinden [2]. Die Anzahl an Lungenkarzinomen nach T0 waren in der CT-Gruppe 297 und in der Röntgen-Gruppe 193 Personen. Außerdem muss berücksichtigt werden, dass in der CT-Gruppe bei fünf Personen und in der Röntgen-Gruppe bei drei Teilnehmern ein Lungenkarzinom vor der ersten Screeningrunde (T0) entdeckt wurde [2, S. 40]. Damit verringert sich die Zahl der Lungenkarzinome, die durch das T0 Screening entdeckt wurden auf 292 (LDCT) und 190 (Röntgen). Darüber hinaus fehlen Angaben im Studienbericht und in den Appendices über die Anzahl an Personen, die das T0 Screening nicht durchgeführt haben oder deren Screeningergebnis

nicht ausgewertet wurde und ein Lungenkarzinom entwickelt haben. Dieses Detail lässt sich lediglich einem weiteren Artikel aus dem Jahre 2013 entnehmen, in welchem die initialen Studienergebnisse des NLST dokumentiert sind und zudem für das T0 Screening eine Grafik abgebildet ist, aus welcher alle Zahlenwerte für die erste Screeningrunde entnommen werden können [16, S. 1983]. Daraus ergibt sich, dass in der CT-Gruppe vier und in der Röntgen-Gruppe fünf Teilnehmer ein Lungenkarzinom bekamen ohne das Screening durchlaufen zu haben. Des Weiteren hatten 57 Patienten mit negativem und 10 mit positiven Ergebnis im CT einen unbekannten Krankheitsstatus. Im Röntgen waren es 8 Teilnehmer mit positivem und 52 mit negativem Resultat. Diese Zahlenwerte müssen von der Gesamtzahl der positiven und negativen Screeningergebnisse abgezogen werden. Für die Screeningrunde T0 lassen sich folgende Vierfeldertafeln erstellen:

Tabelle 11: Vierfeldertafeln für die T0-Screeningrunde

Mit dieser Vierfeldertafel lassen sich Aussagen zur Testgenauigkeit, Sensitivität, Spezifität, positiv und negativ prädiktivem Wert für die T0-Screeningrunde dieser Studie treffen.

(CT: Computertomograph. CT positiv / Röntgen positiv: Der Untersuchungsbefund deutet auf die gesuchte Erkrankung hin. CT negativ / Röntgen negativ: Der Untersuchungsbefund deutet nicht auf die gesuchte Erkrankung hin. Krank: Personen mit der gesuchten Erkrankung, wobei Σ TP+FN die Gesamtzahl der Kranken angibt. Gesund: Personen ohne die gesuchte Erkrankung, wobei Σ FP+TN die Gesamtzahl der Gesunden angibt. TP („True Positive“): Anzahl der Patienten, die nach einem positiven Befund auch tatsächlich an der gesuchten Erkrankung leiden. FP („False Positive“): Anzahl der Patienten, die nach einem positiven Befund nicht an der gesuchten Erkrankung leiden. FN („False Negative“): Anzahl der Personen, die trotz eines negativen Befundes an der gesuchten Erkrankung leiden. TN („True Negative“): Anzahl der Personen, die nach einem negativen Befund nicht an der gesuchten Erkrankung leiden.)

Vierfeldertafel CT

	Krank	***Gesund***	
CT positiv	270 (TP)	6911(FP)	7181
CT negativ	18 (FN)	19043 (TN)	19061
	288	25954	26242

Vierfeldertafel Röntgen

	Krank	***Gesund***	
Röntgen positiv	136 (TP)	2243 (FP)	2379
Röntgen negativ	49 (FN)	23547 (TN)	23596
	185	25790	25975

Aus den Zahlenwerten der oben gezeigten Vierfeldertafeln lassen sich die Sensitivität, Spezifität, der positiv (PPW) und negativ (NPW) prädiktive Wert, die positive (PLHR) und die negative (NLHR) Likelihood Ratio und die Testgenauigkeit berechnen.

Sensitivität (CT) = 270 / 288 = 93,8 %

Spezifität (CT) = 19043 / 25954 = 73,4 %

PPW (CT) = 270 / 7181 = 3,8 %
NPW (CT) = 19043 / 19061 = 99,9 %
PLHR (CT) = 0,938 / (1 – 0,734) = 3,5
NLHR (CT) = (1 – 0,938) / 0,734 = 0,08
Testgenauigkeit (CT) = (270 + 19043) / 26242 = 0,74

Sensitivität (Röntgen) = 136 / 185 = 73,5 %
Spezifität (Röntgen) = 23547 / 25790 = 91,3 %
PPW (Röntgen) = 136 / 2379 = 5,7 %
NPW (Röntgen) = 23547 / 23596 = 99,8 %
PLHR (Röntgen) = 0,735 / (1 – 0,913) = 8,4
NLHR (Röntgen) = (1 – 0,735) / 0,913 = 0,29
Testgenauigkeit (Röntgen) = (136 + 23547) / 25975 = 0,91

Die weiteren Vierfeldertafeln für die Screeningrunden T1 (Tabelle 12) und T2 (Tabelle13) sind im Anhang illustriert. Diese können jedoch aufgrund fehlender Daten zu Patienten, die das Screening nicht durchlaufen haben oder Ergebnisse erzielten, die nicht dichotom zugeteilt werden konnten, nicht korrekt angefertigt werden und beschreiben deshalb nur eine Näherung.

Zusammenfassend lässt sich aus den Berechnungen erkennen, dass das CT-Screening eine gute Sensitivität aufweist und sich daher hervorragend zum Ausschluss einer Erkrankung eignen würde. Dies steht auch im Einklang zum negativ prädiktiven Wert. Es muss allerdings berücksichtigt werden, dass durch die differentielle Verifikation die Sensitivität falsch niedrig angenommen und umgekehrt die Spezifität zu hoch geschätzt wird, da die negativen Screeningergebnisse nicht überprüft wurden [52]. Die PLHR des CTs kommt gerade in einem akzeptablen Bereich (>3) zu liegen, denn ein Wert über 10 würde man als gut bezeichnen. Hingegen liegt die NLHR im aussagekräftigen Bereich (<0,1). Auch diese beiden Werte bestärken die Annahme, dass sich der Test besser zum Ausschluss einer Erkrankung eignet als zu deren Nachweis.

Beim Röntgen-Screening zeigen die berechneten Werte genau in die andere Richtung. Die hohe Spezifität und die gute PLHR lassen den Schluss zu, dass sich das Röntgen für den Nachweis einer Erkrankung eignen würde, da es zudem eine

deutlich höhere Testgenauigkeit aufweist.

Als Fazit muss dennoch für beide Verfahren festgehalten werden, dass sich aufgrund der ungenügenden Zahlenwerte beim PPW kein Verfahren als qualifiziert erweist, eine Erkrankung sicher nachzuweisen. Beide Screeningverfahren liefern hohe Werte an falsch positiven Ergebnissen, die erst durch weitere bildgebende Verfahren oder invasive diagnostische Maßnahmen entdeckt werden können.

(Frage 4.2)

Existieren eindeutige Empfehlungen, ab wann ein Screeningbefund als positiv oder negativ zu bewerten ist?
(Likelihood Ratio Bias)

Erfolgt für einen Screeningtest keine eindeutige Vorgabe ab wann dieser als positiv oder negativ zu bewerten ist, kann dies zu einem Anstieg an falsch positiven und falsch negativen Befunden führen, des Weiteren eine Abhängigkeit von den Fähigkeiten des Untersuchers bedeuten und damit zu einem Likelihood Ratio Bias führen, welcher die interne und externe Validität beeinflusst.

In der Studie vom NEJM wurde berichtet, dass für die Interpretation der CT-Befunde jeder nicht kalzifizierte Knoten, der einen Durchmesser von mindestens 4 mm besaß als positiv eingestuft wurde und zudem verdächtig auf ein Lungenkarzinom war. Bei anderen Abnormalitäten, darunter zählten Adenopathie oder Ergüsse, konnte je nach Ermessen des Radiologen der suspekte Bereich als positiv oder negativ klassifiziert werden. Abnormalitäten, die in T0 (1. Screeningrunde) oder T1 (2. Screeningrunde) als verdächtig eingeordnet wurden und während des Screeningverlaufs konstant geblieben waren, konnten in T2 (3. und zugleich letzte Screeningrunde) als geringfügige Abnormalitäten deklariert werden. Für die Interpretation der Röntgen-Thorax Befunde wurden dieselben Vorgaben gemacht wie beim CT, mit einer Ausnahme: Nicht nur jeder nicht kalzifizierte Knoten über 4 mm Durchmesser wurde als positiv eingestuft, sondern jeder nicht kalzifizierte Knoten oder Masse, unbedeutend welchen Durchmessers, wurde als verdächtig angesehen [1, S. 397].

Nach diesen Vorgaben in der Studie kann nicht von eindeutigen Empfehlungen gesprochen werden, denn die Radiologen konnten nach ihrem Ermessen die

Befunde als positiv oder negativ einstufen. In diesen Kontext fügen sich auch die hohe Anzahl an falsch positiven Befunden ein, denn 96,4% der positiven Screeningergebnisse im CT und 94,5% der positiven Ergebnisse im Röntgen waren falsch positiv [1].

Im Studienprotokoll des NLST [3], ein 924-seitiges Werk mit zusätzlichen Appendices, findet sich eine genauere Definition bezüglich der Zuordnung der Screeningbefunde.

Für einen benignen Knoten im Low-Dose CT sprachen laut dem NLST Protokoll folgende Charakteristiken: Eine Verkalkung, die zentral oder am Rande lag, gleichmäßig verteilt war oder andere Eigenschaften aufwies, die als benigne galten. In den Bereich benigne Läsion fielen ebenfalls Knoten, die eine Fettattenuation oder eine lineare Morphologie besaßen und Veränderungen, die zwei oder mehr Jahre stabil blieben. Knoten, die einen Durchmesser unter 4 mm hatten wurden dokumentiert, jedoch nicht als positiv deklariert [3, S. 29-34].

Als abnormal galten in dem Studienprotokoll des NLST für das Low-Dose CT folgende Merkmale: Jeder neu aufgetretene Knoten über 10 mm Durchmesser oder ein sich in letzter Zeit vergrößernder Knoten über 7 mm, der keine ausreichenden Benignitätskriterien aufwies oder in früheren Untersuchungen nicht eindeutig als gutartig klassifiziert wurde. Die Charakteristiken des Knotens (spikuliert, nicht klar abgrenzbar und andere) wurden zusammen mit seinen Abmessungen dokumentiert [3, S. 29-34].

Im Studienprotokoll wurde noch eine dritte Unterteilung für das CT vorgenommen, die in dem Studienbericht nicht erwähnt wurde. Hierbei handelte es sich um Knoten, die nicht dichotom zugeteilt werden konnten und zu welchem neue solitäre oder multiple Mikroknoten zählten, die einen Durchmesser zwischen 4 und 10 mm oder ein Größenwachstum unter 7 mm besaßen [3, S. 29-34].

Für die Röntgen-Thorax Untersuchungen wurden im Studienprotokoll lediglich zwei Unterteilungen vorgenommen, nämlich in benigne und abnormal. Als benigner Knoten zählten fokale Trübungen, die Verkalkungen zentral oder am Rande trugen und die gleichmäßig verteilte Verkalkungen aufwiesen oder über zwei oder mehr Jahre stabil waren [3, S. 29-34]. Für einen abnormalen Knoten sprachen im Protokoll folgende Details: Ein neu aufgetretener oder sich vergrößernder Knoten, der keine ausreichenden Kriterien für eine Benignität

enthielt oder zeitlich früher als nicht eindeutig gutartig klassifiziert wurde [3, S. 29-34].

Insgesamt gibt das Studienprotokoll im Vergleich zum Studienbericht genauere Leitlinien für die Bewertung eines Screeningergebnisses vor, dennoch bleiben diese Leitlinien ungenau und sind unzureichend. Für die richtige Befundung eines Screeningergebnisses ist die Fähigkeit, Erfahrung und Ausbildung des beurteilenden Radiologen maßgeblich mitentscheidend.

(Frage 4.3)

Unterscheiden sich die teilnehmenden Ärzte in der Studie durch einen höheren Ausbildungsstand von denen außerhalb der Studie?
(Likelihood Ratio Bias)

Die Güte eines Screeningtests kann durch den Ausbildungsstand der Ärzte beeinflusst werden und somit zu einer höheren Trefferquote (Likelihood Ratio Bias) bei der Zuteilung der Screeningergebnisse führen. Dieses Kriterium verringert die Übertragbarkeit in die Allgemeinbevölkerung (externe Validität).

Die Radiologen des NLST bekamen eine Zertifizierung und wurden einer speziellen Unterweisung und Qualitätssicherungsprüfung unterzogen sowie einem Training in der Befundung und Sicherung der Bildqualität [1, S. 397]. In einem 59-seitigen Dokument [5], das zum Overview and Study Design Paper [4] gehört und die Appendices E1-E9 enthält, wurden in Appendix E2 die Zertifizierungskriterien für die Radiologen definiert. Die Zertifizierung der Radiologen erfolgte unter anderem durch das American Board of Radiology. Voraussetzungen waren die Überwachung und Interpretation von mindestens 300 Thorax-CTs in den letzten drei Jahren, zuzüglich mindestens 200 Röntgen-Thoraxe pro Jahr. Zusätzlich wurden regelmäßige Weiterbildungen nach dem American College of Radiology Standard erwartet oder der Nachweis über relevante Anleitungen in der Bildbeurteilung, des Strahlenschutzes, der Bildqualität und den Umgang mit digitalen und nicht digitalen Röntgenbildern im Umfang von 80 Stunden [3, S. 41f.]. Die beteiligten Ärzte erhielten Formulare mit Anweisungen und Beispielfällen für die Interpretation von Befunden und für die Sicherung der Bildqualität. Zudem

gehörten viele der beteiligten Studienzentren zu Zentren, die für ihre ausgezeichneten radiologischen Kliniken bekannt waren [1, S. 405].

Die Frage, ob sich die teilnehmenden Ärzte in der Studie durch einen höheren Ausbildungsstand von denen außerhalb der Studie unterschieden, ist damit eindeutig mit „Ja“ zu beantworten. Die Ärzte, die sich an der Studie beteiligten, mussten eine umfassende Ausbildung nachweisen, unterlagen einem Qualitätssicherungsprogramm und bekamen ein Training, indem sie lernten, wie die CT- und Röntgenbilder befundet und interpretiert werden sollten.

(Frage 4.4)

Hatten die behandelten Ärzte die Möglichkeit, sich klinische Informationen der Patienten anzueignen?
(Likelihood Ratio Bias)

Die Ärzte innerhalb der Studie können durch klinische Informationen über den Gesundheitszustand der Patienten ihre Wahrscheinlichkeit für korrekte Ergebnisse beim Screeningtest künstlich erhöhen und somit das Auftreten eines Likelihood Ratio Bias begünstigen.

Die Radiologen in der Studie hatten Zugriff auf alle klinischen Daten ihrer Patienten, sodass sie Untersuchungs- und Laborergebnisse, Patientenanamnesen und Arztbriefe einsehen konnten, falls diese zur Verfügung standen. Im Studienprotokoll des NLST wurde ausdrücklich darauf hingewiesen, dass die Ärzte zuerst eine Bildbeurteilung ohne Hilfsmittel und im zweiten Schritt einen Vergleich mit Voraufnahmen, unter der Voraussetzung, dass diese vorhanden waren, durchführen sollten [3, S. 28][4, S. 248]. Dass die Radiologen die aktuellen Bilder ihrer Patienten mit früheren Bildern vergleichen sollten, wurde auch im Studienbericht erwähnt und diese Anmerkung verwies damit indirekt darauf, dass die Ärzte Zugriff auf Patientenmaterial hatten [1, S. 397].

(Frage 5)

Existieren Unterschiede in den Rahmenbedingungen, die das Screening, die Diagnostik oder die Therapie betreffen, sowohl zwischen den zu vergleichenden Gruppen als auch zwischen den teilnehmenden Studienzentren?

(Performance Bias)

Beim Performance Bias kommt es zur Inkongruenz zwischen den zu untersuchenden Studienarmen oder zwischen den Studienzentren, wodurch eine Beurteilung der Studienergebnisse erschwert wird (Auswirkungen auf die interne und externe Validität).

Innerhalb der Studie existieren Unterschiede zwischen den teilnehmenden Studienzentren. An 15 von 23 ACRIN-Zentren bekamen die teilnehmenden Patienten zusätzliche Diagnostik in Form von Blutproben, Sputum- und Urinuntersuchungen [1, S. 397]. Das bedeutet, dass fast die Hälfte aller Studienzentren (insgesamt 33) nicht nur das eigentliche Screening durchführten, sondern auch zusätzliche Informationen der Patienten gewannen, die für die Radiologen sichtbar waren. Folglich kann davon ausgegangen werden, dass die Radiologen bei auffälligen Befunden die Bilder genauer auf Anzeichen für Malignität überprüft haben. Des Weiteren ist auch die Übertragbarkeit und Reliabilität auf andere Systeme, wie beispielsweise in Deutschland, gefährdet, da unklar ist, ob die Reduktion der Mortalität nur auf dem Screening oder auch auf den zusätzlichen diagnostischen Maßnahmen beruhte.

Zudem wurden an allen 23 ACRIN-Zentren Daten zur Kosteneffektivität, Lebensqualität und zum Rauchverhalten der Patienten eingeholt [1, S.397]. Diese Patienten bekamen mehr Zuwendung und konnten in Fragebögen über ihre Probleme und Sorgen Auskunft geben. Dies könnte dafür gesorgt haben, dass die Menschen an den ACRIN-Zentren mehr Aufmerksamkeit bekamen und folglich weniger zusätzliche Diagnostik in Anspruch nahmen.

Ein weiterer Unterschied, der zwischen den teilnehmenden Zentren auffällt ist, dass die Befragung der Patienten über ihren Vitalstatus in den LSS-Zentren jährlich und in den ACRIN-Zentren halbjährlich eingeholt wurde [1, S. 398]. Auch hierbei bleibt ungeklärt, ob Symptome bei Patienten in den ACRIN-Zentren früher abgeklärt wurden und so eine zeitlich frühere Diagnose gestellt werden konnte.

Zwischen den beiden Studienarmen gab es ebenfalls Unterschiede bezogen auf die durchgeführten diagnostischen Maßnahmen nach einem positiven Screeningbefund. Im Durchschnitt wurden in der Röntgen-Gruppe mehr invasive Verfahren durchgeführt als in der CT-Gruppe. So wurden etwa 4,8% der

verdächtigen Fälle in der Röntgen-Gruppe im Vergleich zu 4% bei der CT-Gruppe einer chirurgischen Abklärung unterzogen. Mindestens eine zytologische Untersuchung oder Biopsie erhielten 3,5% der positiv befundeten Personen in der Kontrollgruppe und nur 1,8% in der Experimentalgruppe [1, Tabelle 3, S. 401].

Die Behandlung eines Lungenkarzinoms wurde stadienabhängig für die Röntgen- und Kontrollgruppe aufgeschlüsselt [2, Tabelle 3, S. 39]. In dieser Übersicht lassen sich keine bedeutenden Unterschiede zwischen den Gruppen ausmachen, obschon bemängelt werden muss, dass die Behandlung nur stadienabhängig, nicht aber nach histologischem Typ aufgeführt wurde. Die Behandlung eines Lungenkarzinoms richtet sich hauptsächlich nach dem histologischen Typ und sekundär nach dem Stadium, in dem sich das Karzinom befindet [32].

Zusammenfassend kann festgestellt werden, dass auf Grund der bestehenden Unterschiede in den Rahmenbedingungen, die sowohl zwischen den Studienarmen als auch zwischen den Studienzentren auftraten, ein Performance Bias nicht ausgeschlossen werden kann.

(Frage 6)

Sind die Probanden bezüglich ihres Testergebnisses verblindet? Falls nein, besteht die Möglichkeit der zusätzlichen Inanspruchnahme von Diagnostik zur Abklärung eines verdächtigen Befundes außerhalb der Studie? Wie ist das mögliche Problem der Überdiagnose von der gesuchten Erkrankung einzuschätzen?

(Overdiagnosis Bias)

Ein Overdiagnosis Bias kann einerseits auftreten, wenn Studienteilnehmer unnötige diagnostische Maßnahmen außerhalb der Studie in Anspruch nehmen und andererseits, wenn bei Teilnehmern mit einem Lungenkarzinom invasive diagnostische und operative Prozeduren durchgeführt werden, obwohl dieses Karzinom die Patienten in ihrem Leben zu keiner Zeit beeinträchtigt hätte. Im ersten Fall können durch die zusätzlichen diagnostischen Maßnahmen Befunde erzielt werden, die die folgenden Screeningrunden und die daraus entstehenden Ergebnisse beeinflussen können. Im zweiten Fall werden viele Lungenkarzinome detektiert und erfolgreich behandelt und dies zu Gunsten der

Screeningmaßnahme gewertet, obwohl diese Eingriffe unnötig waren.

Die Probanden und die dazugehörigen Hausärzte erhielten innerhalb von vier Wochen nach der Untersuchung einen Brief, indem ihnen ihre Befundergebnisse mitgeteilt wurden [1, S. 397][3, S. 36-37]. Deshalb waren sie bezüglich ihrer Testergebnisse nicht verblindet. Es wurden durch das NLST jährliche Befragungen bei circa 500 Teilnehmern in den LSS-Zentren vorgenommen und damit versucht, die Anzahl an Personen zu ermitteln, die außerhalb der Studie zu einem Lungenkarzinom-Screening gingen. Dabei wurde ein Standardfehler von 0,025 berechnet [1, S. 397], welcher als geringfügig bewertet wurde. Ob diese Befragung in den LSS-Zentren repräsentativ zu werten ist und in welchem Ausmaß wahrheitsgemäß bei den Befragungen geantwortet wurde, bleibt fraglich.

In einem kürzlich erschienenen Artikel wurde ein Versuch unternommen, die Überdiagnose von Lungenkarzinom durch das Screening mit Low-Dose CT im NLST einzuschätzen [59]. Hierbei wurde errechnet, dass die Überdiagnose eines Lungenkarzinoms, welche durch das LDCT-Screening entdeckt werden, bei 18,5% liegt.

Eine weitere Anmerkung wurde von wissenschaftlichen Mitarbeitern der Chiba Universität in Japan 2013 gemacht. Sie beobachteten in mehreren Screeningstudien, unter anderem in dem Italian Lung Trial und in dem NLST, dass die Detektion von Lungenkarzinomen bei der Screeningrunde zu Beginn einer Studie signifikant höher war, als bei den folgenden Screeningrunden. Dies wurde als Hinweis darauf gedeutet, dass indolente Karzinome dazu tendieren zu akkumulieren und so zu einer größeren Rate an Überdiagnosen zu Beginn eines Screenings führen [90]. Dieses Phänomen wurde auch beim Brustkrebs-Screening mittels Mammographie beobachtet [105].

Die Anzahl der Fälle an Überdiagnose im NLST kann nur schwer abgeschätzt werden, da die Nachverfolgungszeit in der Studie äußerst kurz ausgefallen war. Mittels einer längeren Nachbeobachtungszeit könnten die detektierten Lungenkrebsraten mit der Inzidenz in einer Vergleichspopulation ohne Screening gegeneinander abgewogen und so die Anzahl an Überdiagnosen errechnet werden.

(Frage 7)

Wird bei allen Patienten das Screeningergebnis anhand eines Goldstandards verifiziert?

(Verification Bias)

Das Screening beinhaltet den Kasus, dass nur positive Ergebnisse eine weitere Abklärung durch zusätzliche diagnostische Maßnahmen erhalten und damit eine differentielle Verifikation entsteht. Das hat zur Folge, dass nur die falsch positiven Befunde ermittelt werden, nicht jedoch die falsch negativen, die sich lediglich abschätzen lassen (Auswirkungen auf die interne Validität).

Der Goldstandard für den Nachweis eines Lungenkarzinoms ist eine Gewebebiopsie aus dem suspekten Bereich, der durch das Screening entdeckt wurde. Nach einem positiven Screeningbefund wurde hingegen nur bei einem geringen Prozentsatz der Patienten eine Biopsie durchgeführt. Werden chirurgische Maßnahmen zum Goldstandard hinzugezählt, wie eine Mediastinoskopie, Thorakoskopie oder Thorakotomie, dann bekamen etwa 8% der positiv Gescreenten im LDCT und 11,3% im Röntgen eine histologische Gewebeuntersuchung [1, Tabelle 3, S. 401]. Der große Anteil der Personen mit positivem Befund erhielt weiterführende bildgebende Verfahren zur genaueren Abklärung des verdächtigen Herdes (LDCT: 57,9%, Röntgen: 78,4%) [1, Tabelle 3, S. 401]. Das NLST verfasste grob orientierende Richtlinien für das diagnostische Management nach einem positiven Befund [3, S. 28-39]. Die Radiologen innerhalb der Studienzentren konnten das diagnostische Verfahren weitestgehend selbst entscheiden [3, Manual of Operations and Procedures, S. 1-9]. Hierbei muss hervorgehoben werden, dass meistens zwischen dem positiven Screeningbefund und dem Goldstandard (Biopsie/Chirurgische Maßnahmen) ein Zwischenschritt erfolgte, der als eine spezifischere Screeningmaßnahme angesehen werden kann. Dabei handelte es sich unter anderem um eine Verkürzung des Zeitabstandes zwischen zwei Screeninguntersuchungen (von einem Jahr auf 3, 6, 12 oder 24 Monate) oder eine CT-Untersuchung mit der Standard-Strahlendosis oder die Durchführung eines PET-CTs mit FDG [3, S. 29-35]. Viele Patienten mit einem positiven Screeningergebnis bekamen weiterführende bildgebende Verfahren (CT-Gruppe: 57,9% und in der Röntgen-Gruppe: 78,4%) und nur wenige ein

chirurgisches Verfahren (CT-Gruppe: 4% und Röntgen-Gruppe: 4,8%) oder eine Biopsie (CT-Gruppe: < 4% und Röntgen-Gruppe: < 6,5%) [1, Tabelle 3, S. 401].

Die folgende Abbildung veranschaulicht diese Thematik. Die meisten Teilnehmer mit einem suspekten Befund erhielten zuerst weitere bildgebende Verfahren. Erst wenn diese Maßnahmen in die gleiche Richtung wiesen wie der Screeningbefund, wurde ein invasives diagnostisches Verfahren (Goldstandard) durchgeführt, um die Diagnose Lungenkarzinom letztendlich zu sichern und eine Therapie einleiten zu können.

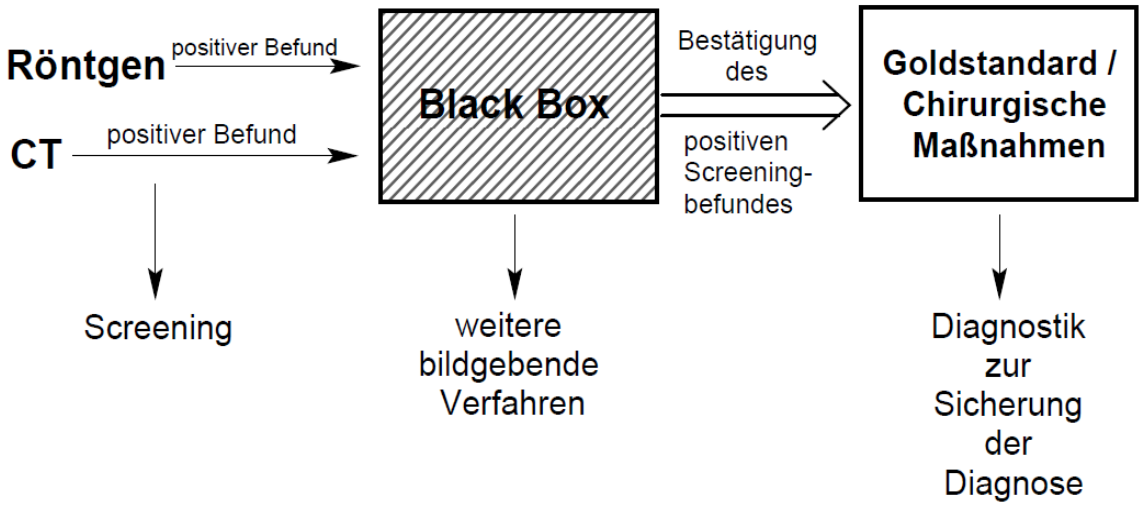

Abbildung 9: Vom positiven Screeningbefund zum Goldstandard

Die obige Darstellung zeigt die Schritte, die bei dieser Studie vom Screening bis zur Sicherung der Diagnose durchlaufen wurden.

Das Screening wurde bei den Teilnehmern der Studie entweder mit Low-Dose CT oder Röntgen durchgeführt. Die sogenannte „Black Box“ beinhaltet eine weiterführende bildgebende Abklärung (ungleich dem Goldstandard) für Patienten nach einem positiven Befund im Röntgen oder CT. Die eigentliche Diagnostik zur Sicherung der Diagnose Lungenkarzinom entspricht einer histologischen Untersuchung des suspekten Gewebes und erfolgt erst im dritten Schritt, wenn das Ergebnis der weiterführenden bildgebenden Verfahren mit dem Ergebnis des Screenings im Einklang stand.

(CT: Computertomograph. Goldstandard: aktuell bester verfügbarer Test zum Nachweis der Erkrankung.)

(Frage 8.1)

Erfolgte eine „Intention to screen“ Analyse?
(Attrition Bias)

Ein Attrition Bias kann in Erscheinung treten, wenn die Studienergebnisse nicht nach der Methode „Intention to screen“ ausgewertet werden. Denn durch die Auswertung von Studienergebnissen von Teilnehmern in der Gruppe, zu der sie nicht ursprünglich randomisiert wurden, kann zum Ungleichgewicht von Personeneigenschaften zwischen den zu untersuchenden Studienarmen führen, obwohl sie zu Beginn der Studie ausgeglichen waren.
In der Studie aus dem NEJM erfolgte eine „Intention to screen“ Analyse, auf die auch ausdrücklich hingewiesen wurde [1, S. 398]. Es wurden alle Ergebnisse der Patienten in der Gruppe ausgewertet, zu der sie auch ursprünglich randomisiert wurden. Bei den Teilnehmern in den LSS-Zentren betrug die jährliche CT-Screeningrate außerhalb der Studie 4,3% in der Röntgen-Gruppe. Die Ergebnisse des CT-Screenings außerhalb der Studie wurde in der Röntgen-Gruppe nach den Maßgaben des „Intention to screen“ ausgewertet [1, S. 399]. Im NLST wurde ein CONSORT Flussdiagramm angefertigt, das den Verlauf der Teilnehmer von der Randomisation bis zur letztendlichen Auswertung zeigt [2, S. 40].

(Frage 8.2)

Sind die Studiengruppen auch noch nach dem Ausscheiden von Studienteilnehmern bezüglich ihrer Personeneigenschaften ähnlich?
(Attrition Bias)

Im NLST Appendix 4 wurde ein CONSORT Flussdiagramm angefertigt, um den Verlauf der Teilnehmer von der Randomisation bis zur finalen Auswertung zu zeigen [2, S. 40]. Es konnte sowohl im Studienbericht wie auch im Studienprotokoll oder den Appendices keine Auflistung der Studienabbrecher nach ihren Personeneigenschaften gefunden werden. In der CT-Gruppe haben bis zur T2 Untersuchung 6,7% (1801 von 26722) und in der Röntgengruppe 10,1% (2697 von 26732) die Studie verlassen oder ein erneutes Screening abgelehnt [2, S.40f.].
In diesem Bereich kann das Auftreten eines Attrition Bias nicht ausgeschlossen

werden [20] und sollte mit in die Überlegungen zur Bewertung der Studie einfließen.

(Frage 9.1)

Sind die Endpunkte der Studie genau definiert? Sind die Dimensionen definiert in welchem die Ziele gemessen werden sollen?
(Detection Bias)

Ein Detection Bias kann entstehen, wenn die Endpunkte nicht eindeutig definiert sind und ein Ermessensspielraum für die Auswertung von Studienergebnissen existiert, sodass zugunsten eines guten Studienoutcomes entschieden werden kann (Auswirkungen auf die interne und externe Validität).
Der primäre Endpunkt versucht herauszufinden, ob durch das Low-Dose CT Screening verglichen mit dem Röntgen-Screening, die Mortalität an Lungenkarzinom (disease-specific mortality) gesenkt werden kann [3, Manual of Operations and Procedures, S. (1-3)]. Der primäre Endpunkt, Tod durch Lungenkarzinom, wurde durch ein Endpoint Verification Process Team, das aus fünf unabhängigen Ärzten bestand, identifiziert [3, Manual of Operations and Procedures, S. (9-2)-(9-3)].
Zum primären Endpunkt wurden gezählt:

1. Totenschein bestätigt Tod durch Lungenkarzinom
2. Tod, der innerhalb von 60 Tagen nach diagnostischen Untersuchungen, die im Zusammenhang mit Verdacht oder der Diagnose Lungenkarzinom stehen, eintrat
3. Tod, der sich während einer therapeutischen Maßnahme bei gesicherter Diagnose ereignete [2, S.33]

Die Zuordnung der Todesursache erfolgte in folgenden Schritten:

1. Der Vorsitzende des Endpoint Verification Teams (EVT) verglich seine identifizierte Todesursache mit der auf dem vorliegenden Totenschein
2. Stimmen beide Todesursachen überein, wird diese so übernommen. Falls nicht, kommt ein zweiter Prüfer hinzu und entscheidet über die Ursache des Todes unter Berücksichtigung der Erstmeinung und der

vorhandenen Informationen über den Todeshergang

3. Bei einer Übereinkunft der Todesursache von Vorsitzendem und zweitem Prüfer wird das Ergebnis übernommen. Falls nicht, erfolgt die Überprüfung durch eine Gruppe von drei EVT Mitgliedern, von denen der Vorsitzende eines der Mitglieder ist [2, S. 33].

Zusätzlich zum primären Endpunkt werden noch sekundäre Studienziele formuliert [3, S. 124][4]. Dazu zählen:

1. Vergleich der Gesamtmortalität (Overall mortality) zwischen dem CT- und Röntgen-Screening [4]
2. Vergleich der Inzidenz [4] und Stadienverteilung des Lungenkarzinoms zwischen den beiden Studienarmen
3. Messung der Testgüte der beiden Screeningverfahren durch Sensitivität, Spezifität, positiv und negativ prädiktiven Werten

Das NLST unterhielt noch einige Substudien in den ACRIN-Zentren, die folgende Ziele verfolgten [3, S. 15]:

- Vergleich des Ressourcenverbrauchs im Zusammenhang mit der Diagnose Lungenkarzinom bei den beiden Studienarmen
- Vergleich der Lebensqualität und der psychologischen Auswirkungen zwischen den beiden Studiengruppen, die das jährliche Screening und positiv Befunde verursachten. Dieser Endpunkt wurde anhand von QALYS (quality-adjusted life–years) gemessen
- Abschätzung der ökonomischen Auswirkungen des Screenings mit CT und Röntgen anhand der Kosteneffektivität
- Etablierung einer Gewebebank von Patienten mit oder ohne gesichertes Lungenkarzinom
- Abschätzung der Auswirkungen des Screenings auf das Rauchverhalten der Teilnehmer. Dies wurde durch die halbjährlich erfolgten Befragungen in den ACRIN-Zentren durch einen Fragebogen vorgenommen [3, S. 23]

Aus der Zusammenschau von Studienprotokoll [3], Appendices [2][5] und dem

Artikel „Overview and Studydesign“ [4] finden sich Definitionen zu Studienzielen. Durch die Zusammenfügung dieser Punkte kann man einen Überblick über die Studienziele, wie oben dargestellt, erreichen.

(Frage 9.2)

Ist die Nachbeobachtungszeit ausreichend lange, um adäquate Endpunkte zu messen?
(Detection Bias)

Eine angemessene Nachbeobachtungszeit ist entscheidend, um Endpunkte korrekt messen zu können. Tatsächliche Auswirkungen einer Screeningmaßnahme können durch die 5-Jahres-Überlebensrate abgeschätzt werden, unter der Berücksichtigung der Existenz eines Lead time bias (Auswirkungen auf die interne Validität).

Das Screening wurde von August 2002 bis September 2007 durchgeführt. Der Tod durch Lungenkarzinom wurde bis Januar 2009 und die Gesamtmortalität der Teilnehmer bis Dezember 2009 ausgewertet [1]. Die Nachbeobachtungszeit für die Teilnehmer betrug längstens 7,4 und mindestens fünf Jahre [3, Manual of Operations and Procedures, S. (1-2)].

Für Männer mit Lungenkarzinom beträgt die relative 5-Jahres-Überlebensrate in Deutschland nur 16% und für Frauen 21%, abhängig vom Krankheitsstadium bei Diagnosestellung [75].

Das NLST schätzt den Lead time bias auf etwa 1,3 Jahre [5], das heißt die Diagnose Lungenkarzinom wird durch das Screening im Schnitt 1,3 Jahre früher gestellt. Unter Berücksichtigung des Lead time bias fällt die Nachbeobachtungszeit von durchschnittlich 6,5 Jahren knapp aus, kann aber in Hinblick auf die niedrigen 5-Jahres-Überlebensraten als ausreichend betrachtet werden. Die PLCO Studie, die sich zeitlich teilweise mit der NLST Studie überschnitt und auch das Screening für Lungenkarzinom untersuchte, wies eine Nachverfolgungszeit von bis zu 13 Jahren auf [55].

(Frage 9.3)

Entsprechen die therapeutischen Komplikationsraten innerhalb der Studie den aktuellen Zahlen außerhalb der Studie?
(Detection Bias)

Eine Studie, die in ausgezeichneten chirurgischen Kliniken durchgeführt wird, kann bessere Therapieergebnisse und folglich ein höheres Studienoutcome aufweisen, als es durchschnittliche Kliniken außerhalb der Studie vermögen und somit Studienergebnisse infolgedessen nicht repliziert werden können (externe Validität).
Die Komplikationsraten, die im Zusammenhang mit therapeutischen Maßnahmen stehen, fallen in der Studie wesentlich geringer aus als sie es außerhalb der Studie sind. Diese Thematik wurde auch in der Diskussion im Studienbericht aufgegriffen und berichtet, dass die Mortalität durch chirurgische Maßnahmen bei circa 1% während des Studienverlaufs lag, verglichen mit durchschnittlich 4% außerhalb der Studie [1, S. 405].
Das Screening wird letztendlich am therapeutischen Erfolg gemessen, denn nur so lassen sich die Mortalität und die Gesamtsterblichkeit an Lungenkarzinom tatsächlich auch senken. Da die durchschnittlichen Komplikationsraten in den Kliniken außerhalb der Studie um das vierfache höher liegen, muss abgeschätzt werden, ob diese Tatsche den Erfolg des CT-Screenings limitiert.

Indirekte Bias

(Frage 10)

Wurde das ideale Studiendesign gewählt?

Für die Bewertung von Screeningverfahren stellen randomisierte, prospektive Parallelgruppen-Vergleiche die Methode der Wahl dar.
Bei der Studie handelte es sich um eine prospektive, kontrollierte, randomisierte, multizentrische Studie im Parallelgruppendesign. Die Feststellung der Todesursachen erfolgte durch das Endpoint Verification Process Team retrospektiv.

Für die Fragestellung wurde das ideale Studiendesign gewählt, jedoch mangelte es an einer adäquaten Umsetzung der Studie.

(Frage 11)

Erfolgte die Allokation zu den Studiengruppen geheim?

Durch eine geheime Allokation können interferierende Placeboeffekte vermieden werden.

Die Allokation erfolgte nicht geheim, denn die Teilnehmer wussten zu jeder Zeit, in welchem Studienarm sie sich befanden. Eine geheime Allokation in dieser Studie wäre aus praktischen Gründen nicht möglich gewesen, da die Teilnehmer die Fragestellung der Studie kannten und die Untersuchungstechnik (CT/Röntgen) leicht für die Teilnehmer zu unterscheiden war.

(Frage 12)

Waren die Ärzte und die Patienten in der Studie verblindet?

Durch eine Verblindung aller Beteiligten kann das Auftreten von verschiedenen Bias ausgeschlossen werden. Darunter fallen zum Beispiel der Information Bias, der Interviewer Bias und der Observer Bias [79].

Die Ärzte konnten nicht verblindet werden, da sie die Untersuchungstechnik anwandten und das Bildmaterial beurteilen mussten. Die Patienten waren weder für das angewendete Screeningverfahren noch in Bezug auf ihre Ergebnisse verblindet.

(Frage 13)

Ist bei Studien im Parallelgruppen-Design der zu vergleichende Indextest akzeptabel?

Ein neues Screeningverfahren sollte nicht mit einem Test verglichen werden, dessen Ineffektivität oder negativen Eigenschaften bereits bekannt sind. Durch deren Vergleich mit solch einem Verfahren steigt die Wahrscheinlichkeit, dass die neue Methode besser abschneiden würde und somit über positive Endergebnisse

der Studie berichtet werden könnte, obwohl das neue Verfahren keinen Benefit gegenüber regulären Untersuchungsmethoden zeigen würde.

Bei den zu vergleichenden Indextests handelte es sich um das Low-Dose CT und die Röntgen-Thorax Untersuchung. Einige Studien, die zeitlich früher durchgeführt wurden als das NLST, verwendeten Röntgen als die Screeningmaßnahme für das Lungenkarzinom [3, S. 6-15]. Auch eine aktuelle Studie, die PLCO (The Prostate, Lung, Colorectal, and Ovarian Randomized Trial), führte das Screening für Lungenkarzinom mit Röntgen durch und verglich die Auswirkungen auf Mortalität und Stadienverteilung der Erkrankung mit Personen, die kein Screening erhalten hatten [55]. Keine der Studien, die Röntgen als Screeningmethode eingesetzt hatten, konnten einen Benefit für die Patienten zeigen.

Da die PLCO Studie sich zeitlich mit der NLST Studie überschnitt, waren zum Zeitpunkt des Studienbeginns des NLST die Ergebnisse des PLCO noch nicht veröffentlicht. Da Ungewissheit darüber herrschte, ob das Röntgen-Screening in dieser Studie einen Nutzen zeigen konnte, wurde das NLST so konzipiert, dass das CT-Screening mit dem Röntgen-Screening verglichen werden konnte [1, S. 405f.].

Es bleibt fraglich, wieso im NLST als vergleichender Indextest das Röntgen eingesetzt wurde, obwohl frühere Studien bereits zeigen konnten, dass das Röntgen-Screening keinen Nutzen für die Patienten erbrachte. Viel Aufwand und Kosten wären erspart geblieben, hätte das NLST eine Experimentalgruppe, die das Low-Dose CT erhielt, mit einer Kontrollgruppe verglichen, die kein Screening bekommen hätte.

(Frage 14)

Können aus der Verteilung der Schweregrade der Erkrankung bei Diagnosestellung Rückschlüsse auf einen Length time bias gezogen werden? Was fällt zusätzlich noch auf?

Bei der Durchführung eines Screenings werden langsam wachsende Tumore eher erkannt als schnell wachsende, da diese lange keine Symptome zeigen (Length time bias).

Im Stadium IA des Lungenkarzinoms befanden sich nach dem ersten Screening

(T0) mit dem Low-Dose CT 45,4% der Patienten, die an einem Lungenkrebs erkrankt sind. Die nachfolgenden Screeningrunden lieferten sogar noch etwas höhere Werte (T1: 47,5%, T2: 50,4%) [2, Tabelle 2]. Fast die Hälfte aller Lungenkarzinome beim CT-Screening wiesen ein Stadium IA auf, sodass ein Length time bias durchaus möglich erscheint. Ein Length time bias besagt, dass beim Screening bevorzugt langsam wachsende Tumore gefunden werden, da die Patienten lange ein symptomfreies Intervall aufweisen können und durch das langsame Voranschreiten der Erkrankung auch ohne Therapie ein längeres Überleben zeigen [52].

Das Röntgen-Screening lieferte keine vergleichbaren Werte, denn hierbei befanden sich nur etwa ein Viertel der Lungenkarzinome im Stadium IA [2].

Die histologische Verteilung der Bronchialkarzinome in der Studienpopulation entsprach nicht dem bekannten und in der Literatur dokumentierten Verteilungsmuster [32][80]. Die nicht kleinzelligen Bronchialkarzinome (NSCLC) machen 85% der Fälle aus, die Kleinzelligen etwa 15% (SCLC) [32]. Nach der WHO Klassifikation von 2004 werden Lungenkarzinome in NSCLC und SCLC eingeteilt [80]. Diese Einteilung erfolgte in der Studie nicht, sondern es wurde teilweise eine Aufgliederung der NSCLC in die histologischen Subtypen vorgenommen [1, Tabelle 5, S. 404].

Allgemein machen unter den NSCLC die Plattenepithelkarzinome mit etwa 40% den größten Anteil aus [32] und zumindest für diesen histologischen Typ wird ein direkter Zusammenhang mit inhalativen Noxen angenommen [80].

Das Plattenepithelkarzinom wird in der Tabelle 5 der Studie nicht aufgeführt [1]. Den größten Anteil nahmen in der Studie die Adenokarzinome mit einem Wert über 35% in beiden Gruppen ein.

Das Adenokarzinom zählt zu den NSCLC und tritt mit einer Häufigkeit von 35% unter den Bronchialkarzinomen auf. Diese histologische Form des Lungenkrebs zählt als die häufigste Karzinomart bei Nichtrauchern und tritt bei Frauen häufiger auf als bei Männern. Als seltene Sonderform des Adenokarzinoms gilt das Bronchioloalveoläre Karzinom [32].

Diese Sonderform wurde gesondert in der Studie aufgeführt und machte in der CT-Gruppe 10,5% aus, was zur Folge hatte, dass sich das aufgetretene Adenokarzinom in der CT-Gruppe auf 46,8% (10,5%+36.3%) summierte. Damit

traten in der CT-Gruppe fast 34% mehr Adenokarzinome auf, als es in der Allgemeinbevölkerung üblich ist.

(Frage 15)

Sind die Nebenwirkungen des neuen Screeningtests für die Patienten akzeptabel?

Der Benefit, der durch die Durchführung einer Screeningmaßnahme erreicht wird, sollte deutlich über dem möglichen Schaden liegen, den Teilnehmer an einer Screeningmethode erleiden können.

Durch das Screeningverfahren (Low-Dose CT) entstehen verschiedene Nachteile für die Patienten.

An denkbaren Nebenwirkungen können unter anderem vorkommen:

1. Folgen durch die Strahlenbelastung
2. Komplikationen durch invasive Diagnostik oder therapeutische Verfahren
3. Psychische Belastungen für die Patienten

Das Low-Dose CT besitzt pro Untersuchung eine Strahlendosis von etwa 1,5mSv, was nur den Bruchteil einer durchschnittlichen CT-Thorax Untersuchung ausmacht (Strahlendosis CT-Thorax: ~8mSv) [9]. Da jeder Patient im experimentellen Studienarm im Schnitt drei Screeningrunden durchlaufen hatte und 39,1% dieser Personen mindestens ein positives Ergebnis hatten, von denen 96,4% falsch Positiv waren [1], bekamen die meisten weitere bildgebende Verfahren [31], um das Screeningergebnis entweder nachzuweisen oder auszuschließen. Hieraus kann berechnet werden, dass sich die pro Kopf Strahlendosis durch das Screening und durch die diagnostischen Maßnahmen auf circa 8mSv summiert und weswegen man zu der Einschätzung gelangen kann, dass eine von 2500 Personen an den Folgen eines Karzinoms, das durch die Strahlenbelastung entstanden ist, sterben wird [9].

Während diagnostischen Verfahren traten bei mindestens 1,4% [(184+61) von (649+17053)] der Teilnehmer in der Low-Dose CT Gruppe Komplikationen auf, wovon 75,1% bei Personen auftraten, die ein Lungenkarzinom hatten. Große Komplikationen traten bei 11,6% der Patienten auf, die an einem diagnostizierten Lungenkarzinom litten, jedoch nur zu 0,1% bei gesunden Teilnehmern. Innerhalb

von 60 Tagen nach einer diagnostischen Maßnahme starben 1,5% der Patienten mit Lungenkarzinom und 0,1% der Gesunden. Aus der Tabelle lässt sich zudem ableiten, dass sechs gesunde Menschen aufgrund von Komplikationen im Zusammenhang mit invasiven diagnostischen Maßnahmen starben, die völlig unnötig waren [1, Tabelle 4, S. 402].

Die Mortalität durch chirurgische Resektionen betrug in der Studie 1% und wird vom NLST außerhalb der Studie auf circa 4% geschätzt [1, S. 405].

Die psychischen Belastungen, die durch dieses Screening für die Betroffenen auftreten konnten, dürfen nicht unterschätzt werden [9]. Zum einen erhielten 39,1% der Patienten ein positives Ergebnis, von denen 96,4% falsch positiv waren [1]. Mehr als jeder dritte Teilnehmer der Experimentalgruppe musste sich mit dem Gedanken auseinandersetzen, dass er möglicherweise an einem Lungenkarzinom erkrankt war, dessen Prognose äußerst ungünstig ist. Die meisten der Patienten mit einem positiven CT-Befund mussten weitere diagnostische Maßnahmen in Kauf nehmen, die schmerzhaft und mit Unannehmlichkeiten verbunden waren, obwohl sie bezogen auf die gesuchte Erkrankung gesund waren. Zum anderen wurden Teilnehmern ein Screeningergebnis mitgeteilt, das zeigte, dass sie sich in einem Indifferenzstadium (z.B.: Knoten im CT, die einen Durchmesser zwischen vier und zehn Millimetern aufwiesen) befanden und zu diesem Zeitpunkt noch nicht eindeutig geklärt werden konnte, ob der Befund positiv oder negativ zu bewerten war. Diesen Personen wurde ein weiteres Screening in einem kürzeren Abstand von drei bis sechs Monaten angeboten, um den Verlauf beobachten zu können [3, S.29-35]. Während dieser Zeitspanne mussten die Beteiligten in Angst und Ungewissheit leben.

Diese drei hervorgehobenen negativen Aspekte des Lungenkarzinom-Screenings mit Low-Dose CT müssen im Verhältnis zur relativen Reduktion der Mortalität an Lungenkarzinom von 20% gesehen und beachtet werden, denn 80% der Menschen mit Lungenkrebs würden zur selben Zeit sterben, unbedeutend ob sie beim Screening waren oder nicht [31].

Tabelle 14: Effekte des Screenings für die Patienten

Die folgende Tabelle soll die entscheidenden positiven und negativen Auswirkungen des Screenings auf die Teilnehmer verdeutlichen.

In der linken Spalte der Tabelle sind die möglichen positiven Auswirkungen des Screenings für die Patienten beschrieben. Auf der rechten Seite der Tabelle sind die möglichen Komplikationen und negativen Auswirkungen des Screenings dargestellt.

Effekte des Screenings für die Patienten	
Positive Effekte	**Negative Effekte**
Frühdiagnose Lungenkarzinom in einem Stadium, indem eine Heilung noch möglich ist	Überdiagnose
Entdeckung von Lungenmetastasen eines bis dato unbekannten Primarius	Strahlenbelastung
Verdeutlichung und Publikmachung des Risikofaktors Rauchen für die Entstehung eines Lungenkarzinoms	Längeres Leben mit der Diagnose Lungenkarzinom ohne Konsequenz für die Verschiebung des Todeszeitpunktes
	Komplikationen durch diagnostische und therapeutische Maßnahmen
Förderung eines gesundheitsbewussten Lebensstils durch die Teilnahme am Screening	Psychische Belastungen durch viele falsch positive Befunde

(Frage 16)

Ist der Screeningtest in der zukünftig zu untersuchenden Gesellschaft ausreichend verfügbar?

Der neue Screeningtest sollte in der Zielpopulation für die der Test konzipiert wurde, auch ausreichend verfügbar sein, damit dieser dort auch erfolgreich implementiert werden kann.

Im Studienprotokoll des NLST existiert eine genaue Auflistung der

Geräteanforderungen, die an das CT gestellt wurden [3, S. 27-28 und Manual of Operations and Procedures, S. (4-2)-(4-3)]. Diese Geräteeigenschaften entsprechen einem durchschnittlichen CT, welches die meisten Kliniken in Deutschland auch besitzen. Deutschlandweit besaßen insgesamt 1001 Krankenhäuser einen Computertomographen im Jahr 2013 [86]. Insgesamt kommen 17,7 Computertomographen auf eine Million Einwohner in Deutschland [85].

Somit könnte die Screeningmaßnahme in Deutschland in nahezu jeder Klinik mit einem CT durchgeführt werden.

Da das Rauchverhalten in den Industrieländern regressiv verläuft und die Entwicklungsländer einen Zuwachs an Rauchern verzeichnen, könnte in naher Zukunft auch hier das LDCT als Screeningmaßnahme an Bedeutung gewinnen. Hierbei sollte allerdings an limitierenden Faktoren beachtet werden, das diese Länder sich ein solch teures Screening nicht leisten können, CTs in Entwicklungsländern kaum verfügbar sind und die hohe Durchseuchungsrate mit Tuberkulose die Zahl an falsch positiven Befunden in die Höhe schnellen lassen würde, wodurch die Kosten für das Screening nochmals ansteigen würden.

(Frage 17)

Bestehen effektive Therapiemöglichkeiten für die gescreente Erkrankung?

Der Erfolg einer Screeningmaßnahme hängt maßgeblich davon ab, ob die gesuchte Erkrankung auch erfolgreich therapiert werden kann. Sind bereits die Chancen auf eine Kuration der Erkrankung im Frühstadium bescheiden, bringt die durchgeführte Screeningmethode mehr Nachteile (z.B. längeres leben mit der Diagnose ohne den Zeitpunkt des Todes durch die Erkrankung beeinflussen zu können) für die Patienten als Vorteile.

Die Therapie des Bronchialkarzinoms erfolgt stadienabhängig und nach der histologischen Zuordnung des Tumors. Das kleinzellige Bronchialkarzinom tendiert dazu, sehr früh hämatogen und lymphogen zu metastasieren. Therapie der Wahl in den Stadien I-III ist die kombinierte Radiochemotherapie [29]. Die 2-Jahres-Überlebensrate beträgt beim SCLC im Stadium „Limited disease“ 50% und im Stadium „Extensive disease“ unter 20% [80]. Für das SCLC besteht bereits im

Stadium I nur eine einschränkte Prognose.

Die nicht kleinzelligen Lungenkarzinome beschränken sich lange auf ein lokales Wachstum, bevor Fernmetastasen auftreten. Aus diesem Grund ist bei frühen Stadien die Operation das Mittel der Wahl [80]. Die 5-Jahres-Überlebensrate beträgt bei den NSCLC im Stadium IA 50% und im Stadium IV 2% [29]. Die NSCLC weisen im Vergleich zu den SCLC die bessere Prognose auf und profitieren durch eine frühe Diagnose (chirurgische R0-Resektion möglich). Im Stadium IA liegt die postoperative 5-Jahres-Überlebensrate bei 55-80% und im Stadium IB bei 55-60% [56]. Im Vergleich zu anderen Karzinomerkrankungen, wie beispielsweise dem Mammakarzinom (5-Jahres-Überlebensrate im Stadium I 100% und im Stadium II 93%) [6] weist das Lungenkarzinom eine eingeschränkte Prognose auf.

(Frage 18)

Bestehen Interessenskonflikte?

Bestehen Interessen eines Investors in einem positiven Studienoutcome, können Studienergebnisse in die gewünschte Richtung interpretiert werden.

Es kann davon ausgegangen werden, dass ein möglicher Interessenskonflikt vorliegt, denn die Studie wurde unter anderem von dem American College of Radiology Imaging Network (ACRIN) verwaltet [1, S. 396] und auch in 33 dieser Zentren zu einem großen Teil durchgeführt. Die Implementierung dieser Screeningmethode würde für die Radiologen einen enormen Patientenzuwachs und folglich eine Umsatzerhöhung bedeuten. Diese Vermutung wird dadurch unterstützt, dass Gruppen, die einen direkten positiven Effekt durch die Einführung dieser Screeningmaßnahme erwarten, das Screening eher empfehlen, wie zum Beispiel die American Association for Thoracic Surgery und die American Lung Association, als Vereinigungen, die davon nicht direkt betroffen sind und einen objektiveren Blickwinkel besitzen (z.B.: American Cancer Society, American Academy of Family Physicians) [19].

(Frage 19)

Ist die Statistik angemessen?

In den dargestellten Statistiken aus dem Studienbericht des NEJM konnten keine groben Fehler gefunden werden. Da jedoch bereits die Existenz einiger Bias aufgezeigt werden konnte, ist es an dieser Stelle von untergeordneter Bedeutung, ob die Statistik Fehler enthält, denn die verwendeten Studienergebnisse der Statistik müssen angezweifelt werden.

Tabelle 15: Zusammenfassende Beurteilung der Studie

In der folgenden Tabelle werden die positiven und negativen Aspekte der untersuchten Studie aufgezeigt. Auf der linken Seite der Tabelle sind die positiven Aspekte der Studie abgebildet. Auf der rechten Seite der Tabelle sind die negativen Aspekte der Studie illustriert. Die einzelnen Punkte wurden in diesem Kapitel 3.3 bereits beschrieben und die Auswirkungen auf die Studie erläutert. Diese Tabelle dient als tabellarische Zusammenfassung aller wichtigen Studiendetails.
(Intention to screen: Auswertung der Studienergebnisse der Probanden in der Gruppe zu der sie ursprünglich randomisiert wurden. CONSORT: Consolidated Standards of Reporting Trials. ACRIN: American College of Radiology Imaging Network. Length time bias: Screening entdeckt bevorzugt langsam wachsende Tumore, da diese längere Zeit symptomfrei bleiben.)

Zusammenfassende Beurteilung der Studie	
Positive Aspekte der Studie	**Negative Aspekte der Studie**
Ausreichend große Studienpopulation	Studienteilnehmer wurden durch die Setzung von Anreizen geworben
Geeignete Wahl der Hochrisikogruppe	Studienarme sind bereits zu Beginn der Studie nicht vergleichbar (mehr T IV Stadien in der Kontrollgruppe beim T0 Screening)
Wahl des idealen Studiendesigns	Hohe Rate an falsch positiven Ergebnissen
Durchführung einer Intention-to-screen Analyse	Ungenaue Empfehlungen wie ein radiologischer Befund zu bewerten ist
Berücksichtigung vieler relevanter Personeneigenschaften	Unterschiede im Ausbildungsstand der Ärzte inner- und außerhalb der Studie

Angemessene Statistik	Ärzte hatten Zugriff auf Patienteninformationen
	Durchführung weiterer Untersuchungen (Blut-, Sputum- und Urinuntersuchungen) und Durchführung zusätzlicher Befragungen in den ACRIN-Zentren
Umfangreiches Studienprotokoll	Überdiagnose Lungenkarzinom durch das Screening
	Keine Verblindung der Studienteilnehmer und Ärzte
Plausible Einschlusskriterien für die Aufnahme der Studienprobanden	Keine geheime Allokation zu den Studienarmen
	Knapp bemessene Nachbeobachtungszeit von durchschnittlich 6,5 Jahren
Eindeutige Definitionen der Studienziele und Endpunkte	Mögliches Auftreten eines Length time bias
	Keine gesonderte Aufschlüsselung nach dem Plattenepithelkarzinom für das der Nachweis für die Entstehung eines Lungenkarzinoms durch inhalative Noxen gesichert ist
Verwendung eines CONSORT Flussdiagramms für die Illustrierung der Screeningrunden T0 bis T2	Keine Durchführung einer Präferenzanalyse der Studienteilnehmer
Ausreichende Verfügbarkeit der Screeningmaßnahme in den Industrieländern	Keine Aufschlüsselung der Studienausscheider nach ihren Personeneigenschaften
Reflektierende Diskussion der Studienergebnisse am Ende des Studienberichts	Der neue Screeningtest eignet sich besser für den Ausschluss einer Krankheit als für deren Nachweis
Veröffentlichung der initialen Studienergebnisse nach dem T0 Screening	Kein Ausschluss möglicher Interessenskonflikte
	Eingeschränkte Prognose für das Lungenkarzinom im Frühstadium
Publizierung der Kosteneffektivitätsanalyse	Durchführung an renommierten radiologischen Zentren

Zusammenfassend kann festgestellt werden, dass Hinweise auf einen Incentive Bias, Selection Bias, Likelihood Ratio Bias, Performance Bias, Overdiagnosis Bias, Verification Bias und Length time bias bestehen. Zudem können Interessenskonflikte innerhalb der Studie und die Existenz des Attrition Bias nicht gänzlich ausgeschlossen werden.

Aufgrund dieser Reihe an Hinweisen und des gerade eben noch akzeptablen Wertes beim PLHR des Low-Dose CT-Screenings, kann das Fazit gezogen werden, dass man dem neuen Screeningtest kritisch gegenüber stehen sollte und Ergebnisse in Kliniken, die das Low-Dose CT-Screening für Raucher bereits eingeführt haben, überprüft werden müssen, bevor eine bundesweite Empfehlung für die Implementierung dieses Screenings ausgeschrieben werden kann.

4. Diskussion

4.1 Unterschiede zwischen den aufgezeigten Mitteln zur Beurteilung von Studien und Vergleich mit der vorgeschlagenen Evaluationsmethode

Im Kapitel Material und Methoden wurden die häufig zitierten Evaluationsmethoden aufgezeigt, welche nun voneinander abgegrenzt und mit der vorgeschlagenen Methode in dieser Arbeit verglichen werden sollen.

Mit dem CONSORT Statement wurde 1996 erstmals eine Leitlinie für Therapiestudien eingeführt, die eine Empfehlung für Autoren, Editoren und Lesern darstellen sollte, um eine vollständige Berichterstattung von Studien zu ermöglichen [83]. Bei der Analyse der CONSORT Checkliste, die 25 Themen abhandelt, wird ersichtlich, dass diese Checkliste allein für die Ersteller eines Studienberichts konzipiert wurde und nicht für den Leser einer Studie. Die Themen, die in einer Therapiestudie eine wichtige Stellung besitzen und über die berichtet werden sollte, werden mit Aufforderungen an die Berichterstatter einer Studie versehen, z.B. wird beim Thema 18 (Zusätzliche Analysen) in der Checkliste angegeben: „Resultate von weiteren Analysen, einschließlich Subgruppenanalysen und adjustierten Analysen mit Angabe, ob diese präspezifiziert oder exploratorisch durchgeführt wurden“ [83, CONSORT Checkliste, Frage 18]. Hierbei ist zudem erkennbar, dass dieser Stoff nur von einer Person beantwortet werden kann, welche die Studie und deren Durchführung sehr gut kennt. Des Weiteren wird ein durchschnittlicher Leser einer Studie (z.B. ein Mediziner), welcher nicht im Besitz von spezifischen Kenntnissen in Studienanalysen ist, kaum in der Lage sein, zu unterscheiden, ob eine Subanalyse präspezifiziert oder exploratorisch durchgeführt wurde. Das dieses Instrument allein für Therapiestudien entwickelt wurde, zeigt sich auch in dem dazugehörigen Flussdiagramm, denn dort werden lediglich die beiden zu vergleichenden Interventionen (Therapien) und die folgende Nachbeobachtungszeit für die

Teilnehmer untersucht. Bei diagnostischen Studien erfolgt noch ein Zwischenschritt, denn die Ergebnisse, die in den zu vergleichenden Diagnostiktests ermittelt wurden, müssen in einem zweiten Schritt durch einen Referenztest (Goldstandard) verifiziert werden (vergleiche dazu das STARD Flussdiagramm [13], Abbildung 2).

Mit STARD entstand 2003 eine Leitlinie für Diagnostikstudien, die genauso wie CONSORT angibt, für Autoren, Editoren und Leser konzipiert worden zu sein. Als primäres Ziel steht die Verbesserung des Studienreports im Mittelpunkt und als sekundäres Ziel soll den Lesern einer Studie ein Hilfsmittel gegeben werden, um potentielle Bias in Studienberichten zu identifizieren [13]. Die STARD Checkliste kann für Leser einer Diagnostikstudie angewendet werden und überzeugt durch kurze und detaillierte Formulierungen verschiedener Themen, die auch Leser einer Studie beantworten können. Im Unterschied zur STARD Checkliste werden indem entwickelten Fragebogen zur Evaluierung von Screeningstudien in dieser Arbeit noch Punkte abgehandelt, die vor allem für Studien zur sekundären Prävention relevant sind und für diagnostische Studien eine kleinere Gewichtung einnehmen. Dies sind beispielsweise Fragen zum Selection Bias (Frage 3.2), Overdiagnosis Bias (Frage 6), Verification Bias (Frage 7), Attrition Bias (Fragen 8.1 und 8.2), Detection Bias (Frage 9.2) und zusätzlich noch die Fragen zu den Indirekten Bias (Fragen 13, 14, 16, 17).

Der 14 Themen umfassende Fragebogen von QUADAS aus dem Jahr 2003 wurde ebenfalls für Berichterstatter von Diagnostikstudien entwickelt und versucht die Qualität von Studien zu verbessern [101]. Der Schwerpunkt des Fragebogens liegt auf der Beurteilung der Qualität, Durchführung und Tauglichkeit des Index- und Referenztests, denn darauf beziehen sich neun der 14 Fragen und im Vergleich zu STARD rückt die Überprüfung der externen Validität in den Hintergrund. Die Patientenrekrutierung, die Randomisation der Studienteilnehmer, die Personeneigenschaften und die Vergleichbarkeit der Studiengruppen werden im QUADAS Fragebogen nicht berücksichtigt.

Ein Artikel, der sich ausdrücklich auch auf die Beurteilung von Screeningstudien bezieht entstand 2006 und hob drei wichtige Aspekte hervor, die elementar sind um eine Testevaluation durchführen zu können [24]. Diese Thesen, die Bestimmung der Testgenauigkeit, die Bedeutung des Tests für die weiterführende

Behandlung und die Bedeutung für das Patienten-Outcome wurden auch in meine Überlegungen zum Screeningablauf mitaufgenommen und in den Fragen zum Likelihood Ratio Bias, Detection Bias und bei den indirekten Bias in den Fragen 14 und 17 verankert.

Aus der Zusammenschau der bereits existenten Mittel zur Evaluierung von Studien, lässt sich schlussfolgern, dass gute Methoden zur Studienbeurteilung vorhanden sind, jedoch wird noch zu wenig Aufmerksamkeit darauf gelegt, diagnostische von Screeningstudien zu unterscheiden. Das diese beiden Studientypen unterschieden werden sollten, wurde von mir im Kapitel 2.3 dargelegt. STARD bietet für Leser einer Studie ein gutes Werkzeug, um eine Studienanalyse durchführen zu können und könnte auch für Screeningstudien zunächst einmal angewendet werden. Treten bei einer Screeningstudie durch die Anwendung der STARD Checkliste Ungereimtheiten auf, sollte das zu untersuchende Testverfahren in Bezug auf dessen Tauglichkeit hinterfragt werden. Tritt jedoch die gegensätzliche Situation auf, dass durch die STARD Checkliste keine Fragen offen bleiben und der Screeningtest in Ordnung zu sein scheint, dann sollte berücksichtigt werden, dass einige Bias und andere mögliche Fehlerquellen mit dieser Methode ignoriert werden (siehe dazu den Vergleich zu STARD mit meinem Fragebogen auf S. 85), da dieses Verfahren primär für diagnostische Studien und nicht für Screeningstudien konzipiert wurde.

4.2 Bedarf es einer Erweiterung des Ausbildungscurriculums für Medizinstudenten in Bezug auf die Bewertung von Screeningmaßnahmen?

In der Approbationsordnung für Medizinstudenten wird darauf verwiesen, dass das Thema Prävention gelehrt und später im Staatsexamen auch abgeprüft werden soll [8]. Doch es wird heutzutage zu wenig Wert darauf gelegt, auf die Bewertung von Screeningmaßnahmen einzugehen. Es wird in den Medien viel über den Nutzen einer Screeningmaßnahme gesprochen, doch in den untersuchten

Lehrbüchern verweist lediglich eine Tabelle im Werk „Harrisons Innere Medizin“ auf die Steigerung der durchschnittlichen Lebenserwartung, die durch ein bestimmtes Screening erreicht werden kann [44]. Der Begriff Prävention wird in vielen Lehrbüchern definiert [82][28][88], doch über den tatsächlichen Nutzen und die möglichen Nachteile wird kaum berichtet. Screeningprogramme sind in der heutigen Zeit ein weit verbreitetes Instrument und werden von vielen Patienten in Anspruch genommen, sodass die Ärzte über ein ausreichendes Wissen und Informationen über die angewendeten Präventionsmaßnahmen verfügen sollten, um Aufklärungen adäquat durchführen zu können. Aus diesem Grund bin ich der Überzeugung, dass bereits im Studium der Humanmedizin die Grundlage dafür geschaffen werden sollte, Präventionsprogramme reflektierend zu betrachten und eigenständige Entscheidungen über die Bewertung von Screeningmaßnahmen zu fördern und dass dies bisher nur unzulänglich geschehen ist.

4.3 Evaluationsmethode für Screeningstudien

Für die Entwicklung einer Evaluationsmethode für Screeningstudien wurde eine aktuelle Studie mit dem Titel „Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening“ aus dem NEJM verwendet [1]. Bis heute existiert kein von den Krankenkassen anerkanntes Screeningverfahren für die Früherkennung von Lungenkrebs in Deutschland. Es wurden bereits mehrere Studien mit dem Ziel durchgeführt, einen verlässlichen Test für die Früherkennung von Lungenkarzinom zu entdecken, jedoch bis zur Durchführung des NLST ohne Erfolg [92][45][89][36][100].

In Kapitel 3.2 wurden diejenigen Bias aufgeführt, welche in der NLST Studie identifiziert wurden. Zusätzlich zu diesen Biasformen existieren noch weitere, wie zum Beispiel der Information Bias, Contamination Bias, Differential Expertise Bias oder der Observer Bias. Aus Gründen der Übersichtlichkeit wurden mehrere Bias zusammengefasst, während andere gänzlich weggelassen wurden, da sie lediglich eine eingeschränkte Relevanz bei Screeningstudien aufweisen.

Der Information Bias tritt durch „(...) Kenntnis der bisherigen Diagnostikbefunde (…)“ auf, welcher „(...) die Interpretation eines weiteren Tests“ beeinflusst [79, S. 231]. Diese Biasform wurde beispielsweise unter dem Likelihood Ratio Bias zusammengefasst (Tabelle 3.4, Frage 4.4), da klinische Informationen zu bereits durchgeführten Untersuchungen die Trefferquote für richtige Befunde erhöht und somit die Likelihood Ratio beeinflusst. Der Differential Expertise Bias wurde ebenfalls unter den Likelihood Ratio Bias subsumiert (Tabelle 3.4, Frage 4.3). Dieser Bias berücksichtigt den Ausbildungsstand der Ärzte in der Studie in Bezug auf bestimmte Untersuchungs- oder Therapietechniken [79]. Durch eine bessere Ausbildung können Unterschiede zwischen Kontroll- und Experimentalgruppe auftreten, diese verzerrt darstellen und eine Übertragbarkeit in andere Kliniken verringern.

Der Contamination Bias fällt in meiner Arbeit mit dem Overdiagnosis Bias zusammen, da sich durch die zusätzliche Inanspruchnahme von Diagnostik vor allem der Kontrollgruppe außerhalb der Studie die Überlegenheit einer Maßnahme gegenüber einer anderen nicht mehr deutlich darstellen lässt [79]. Der Contamination Bias lässt sich nur schwer vom Overdiagnosis Bias abgrenzen. Die Problematik des Contamination Bias kann erkannt werden, wenn eine Auswertung nach der Intention-to-screen Methode mit einer Auswertung nach der Per-Protocol-Analyse (nur die Studienergebnisse der Teilnehmer werden ausgewertet, die sich ans Protokoll gehalten haben und nicht frühzeitig ausgeschieden sind) verglichen wird [79]. Der Observer Bias fand in dieser Dissertation zum Beispiel keine direkte Berücksichtigung, da es für den Leser einer Studie äußert schwierig wird zu entscheiden, ob Ergebnisse in einer Studie zu Gunsten der Hypothese interpretiert wurden. Dennoch wurde dieser Bias indirekt erfasst, indem nach der Verblindung der Ärzte und Studienteilnehmer gefragt wurde, denn damit kann dieser Effekt zumindest minimiert werden. Zur Verwirklichung eines übersichtlichen und einfach zu handhabbaren Instruments für Leser einer Screeningstudie wurde auf eine vollständige Aufschlüsselung aller denkbaren Biasformen für die Evaluierung verzichtet. Demzufolge besteht allerdings die Gefahr, dass die Existenz einiger Biasarten in einer Studie nicht entdeckt werden und dadurch einem Test eine höhere Qualität und Validität zugesprochen wird als ihm eigentlich zusteht. Ein weiterer Punkt den es zu beachten gilt, ist der, dass

den verwendeten Biasformen konkrete Fragen zugeteilt wurden, die vom Leser einer Studie einfach und unmissverständlich beantwortbar sein müssen. Mit dieser Methode wird allerdings die Aussagekraft der entwickelten Fragensammlung eingeschränkt, denn durch die Verwendung spezifischer Fragen werden markante Beispiele für einen Bias auf Kosten der vollständigen Überprüfung jeder einzelnen Biasform herausgegriffen. Bei den indirekten Bias wurde wiederum eine Auswahl aus dem erweiterten USP-Fragebogen (Tabelle 7) getroffen, die Sachverhalte zu berücksichtigen versuchen, die eine Studie betreffen. Hierbei wurde das Augenmerk vor allem darauf gelegt, nur solche Tatbestände zu erfassen, die direkte Folgen für die Validität und klinische Relevanz einer Studie haben. Folglich wurden Punkte nicht berücksichtigt, die sich mit dem Informed Consent oder der Möglichkeit der Widerrufung der Teilnahme beschäftigen. Des Weiteren wurden die Zeitintervalle zwischen den einzelnen Tests und zwischen Test und Therapie nicht berücksichtigt, da in diesem Fall ein direkter Rückschluss auf die Validität eines Testverfahrens nicht möglich ist.

Ein weiterer wichtiger Punkt, der die erfolgreiche Anwendung dieses Verfahrens minimiert ist, dass immer noch Handlungsbedarf besteht, damit RCTs korrekt und vollständig berichtet werden. Dies zeigt die Dissertation von Marc Nagel aus dem Jahr 2007. Hierbei konnte ermittelt werden, dass die Intervention wie sie im Titel einer Studie benannt wurde, nur bei 49 von 53 Studien auch im Methodenteil durchgeführt wurde. Außerdem stimmte das Ergebnis über das in der Diskussion berichtet wurde, nur bei 42 von 53 Studien mit dem vermuteten Ergebnis aus dem Titel überein [51]. Diese Arbeit lässt die Schlussfolgerung zu, dass noch Mängel in der Studienberichterstattung bestehen, die es für den Leser einer Studie erschweren, über die Validität eines neuen Tests oder einer Therapie eine Aussage zu treffen.

4.4 Interpretation der Studienergebnisse des NLST

Durch die Anwendung der vorgestellten Fragensammlung im Ergebnisteil Kapitel 3.2 konnten einige Aspekte der Studie zum Vorschein gebracht werden, die der näheren Betrachtung und Diskussion bedürfen. Diese Themen werden im Folgenden analysiert.

Der Healthy volunteer Bias wird als Untergruppe des Incentive Bias geführt und durch diesen Bias werden bevorzugt gesündere Personen in eine Studie eingeschlossen. Es erhebt sich die Frage, ob dies für eine Studie zur sekundären Prävention günstig ist.

Negative Folgen sind zum einen, dass durch den Healthy volunteer Bias ein Sampling Bias entsteht, da die Studienpopulation sich von der späteren Zielbevölkerung unterscheidet und somit die Reproduzierbarkeit (externe Validität) von Studienergebnissen eingeschränkt wird. Zum anderen kann die Prävalenz von Lungenkarzinomen bei den Studienteilnehmern einen kleineren Wert annehmen und damit provozieren, dass der Screeningtest mehr falsch positive Ergebnisse liefert und somit der positiv prädiktive Wert sinkt [64]. An positiven Effekten sind zu nennen, dass Personen, die zum Zeitpunkt der Diagnose Lungenkrebs nicht mehr rauchen oder daraufhin das Rauchen aufgeben, ein signifikant besseres Therapieansprechen und eine verbesserte Lebensqualität aufweisen, als jene, die das Rauchen nicht aufhören [7]. Im Detail konnte nachgewiesen werden, dass Nichtraucher bei Diagnosestellung weniger peri- oder postoperative Komplikationen hatten, ein besseres Ansprechen auf Chemotherapien zeigten, weniger Zweitkarzinome entwickelten und eine geringere Gesamtsterblichkeit zeigten (Gesamtmortalität ist bei Rauchern 2.94 mal höher). Beim Healthy volunteer Bias weisen die Personen einen gesundheitsbewussten Lebensstil auf und haben damit eher die Tendenz das Rauchen zu beenden, wenn dadurch ein Nutzen für ihre Genesung erzielt werden kann. Eine Möglichkeit den healthy volunteer effect nachzuweisen besteht darin, die Studienpopulation auch in Bezug auf das Auftreten anderer Karzinomerkrankungen zu untersuchen. Treten bei den

Teilnehmern in der Studie insgesamt weniger Krebserkrankungen auf als es durchschnittlich zu erwarten gewesen wäre, liegt der Verdacht nahe, dass ein Healthy volunteer Bias besteht.

Bei dem Punkt Risk identification Bias wird bemängelt, dass in der NLST Studie ein Alter zwischen 55 und 74 Jahren als Einschlusskriterium gewählt wurde. Die Neuerkrankungsrate an Lungenkarzinom erreicht ihre höchsten Werte jedoch erst nach dem 74. Lebensjahr. Diese Patienten wurden aber aus der Studie ausgeschlossen. Es muss dennoch berücksichtigt werden, dass das Lungenkarzinom-Screening lediglich eine Früherkennung einer Karzinomerkrankung ist und die Erkrankung als Präventivmaßnahme nicht verhindert. Folglich ist dieses Screening nur dann sinnvoll, wenn eine Therapie mit kurativer Intention eingeleitet werden kann und dies stellt in diesem Fall die Operation dar. Für Patienten jenseits des 74. Lebensjahres ist jedoch die ethische Rechtfertigung einer kurativen operativen Maßnahme fraglich, da die Wahrscheinlichkeit an einer anderen Erkrankung als dem Lungenkarzinom zu sterben hoch ist. Deshalb ist die gewählte Altersspanne für die Studie durchaus berechtigt und nachvollziehbar. Die anderen Risikofaktoren wurden korrekt gewählt. Dennoch bleibt erwähnenswert, dass mindestens 15 bis 20% aller Personen, die ein Lungenkarzinom entwickeln werden, nicht in dieses Risikoprofil passen und somit nie zu einem Screening geladen würden [58].

Darüber hinaus besteht der Verdacht, dass die untersuchte Studie einen Selection Bias aufweist, da die Röntgen-Gruppe im Vergleich zur CT-Gruppe bereits beim T0 Screening mehr Patienten mit einem Stadium IV eines Lungenkarzinoms aufwies [2, Tabelle 2]. Diese Verteilung ist weitestgehend unabhängig von der durchgeführten Screeningmaßnahme, da das T0 Screening direkt nach der Randomisation durchgeführt wurde und im Röntgen-Screening weniger Knoten entdeckt werden können, da das Röntgen eine schlechtere Auflösung als das LDCT besitzt. Aus diesem Grund muss davon ausgegangen werden, dass in der CT-Gruppe weniger Teilnehmer ein Stadium IV eines Lungenkarzinoms hatten, denn diese wären im LDCT eher entdeckt worden. Diese Stadienverteilung zu Beginn der Studie führt zu der Annahme, dass die beiden Studienarme trotz der ähnlichen Personeneigenschaften, die in der Tabelle 1 des Studienberichts von 2011 abgebildet wurden, nicht vergleichbar waren und sich in Eigenschaften

unterschieden, die in der Studie nicht bestimmt wurden. Dies führt dazu, dass die Röntgen-Gruppe schlechtere Startbedingungen hatte als die LDCT-Gruppe und somit die Überlegenheit des LDCT-Screenings relativiert werden muss, da hierdurch die Studienergebnisse zu Gunsten des LDCTs verzerrt dargestellt wurden.

Zudem muss die Existenz eines Likelihood Ratio Bias angenommen werden, denn alle drei Sachverhalte, die in diesem Zusammenhang im Fragebogen abgefragt wurden, deuten in diese Richtung. Durch die Berechnungen, die bei Frage 4.1 vorgenommen wurden, konnte gezeigt werden, dass sich der LDCT-Screeningtest eher für den Ausschluss als für den Nachweis einer Erkrankung eignet, da der Screeningtest eine ausgesprochen hohe Sensitivität aufweist. Ferner wurden keine eindeutigen Empfehlungen für die Bewertung eines Befundes als positiv oder negativ vorgelegt, sondern eher vage Empfehlungen ausgesprochen, bei denen die Radiologen eine große Entscheidungsfreiheit hatten. Zusätzlich wurde den Ärzten innerhalb der Studie eine spezielle Ausbildung für die korrekte Interpretation von Befunden zuteil. Als vierter entscheidender Punkt ist zu nennen, dass die Radiologen Zugriff auf klinische Informationen zu ihren Patienten hatten und mit diesem Hintergrundwissen ihre Trefferquote für richtige Befunde zusätzlich erhöhen konnten.

Diese verschiedenen Aspekte des Likelihood Ratio Bias führten dazu, dass die erzielten Ergebnisse der Screeningstudie nicht nur allein dem neuen Screeningtest, sondern zu einem erheblichen Anteil den Radiologen zugeschrieben werden konnten, deren Fähigkeiten durch ihre erweiterte Ausbildung, durch die Durchführung der Studie an renommierten radiologischen Zentren und durch klinischen Informationen, die sie bezüglich ihrer Patienten einholen konnten und sogar sollten, verbessert wurden. Diese Aspekte führten dazu, dass der Screeningtest besser abschnitt, als er es unter realen Bedingungen tun würde. Zudem muss bei den aufgeführten Punkten des Likelihood Ratio Bias noch angemerkt werden, dass hier zusätzlich noch ein Performance Bias aufgetreten war, da die Zuteilung der Befunde im LDCT zu positiv, negativ oder nicht eindeutig feststellbar (Indifferenzstadium) erfolgte, während im Röntgen die Zuteilung nur zu positiv oder negativ möglich war. Dieser Unterschied in der Auswertung von Bildbefunden führte zu einem Performance Bias.

Weitere Ungleichheiten bestanden zwischen den teilnehmenden Studienzentren. Die ACRIN-Zentren führten zusätzliche Untersuchungen in Form von Sputum-, Blut- und Urinanalysen durch und nahmen Befragungen der Teilnehmer vor, die in den LSS-Zentren nicht stattfanden. Diese verschiedenen Maßnahmen, die in den Studienzentren vollzogen wurden, führten letztendlich dazu, dass den Teilnehmern an den ACRIN-Zentren mehr Aufmerksamkeit gewidmet wurde und somit sich Studienteilnehmer eher Regelkonform ans Protokoll hielten, als in den LSS-Zentren.

Die Patienten waren zu keiner Zeit verblindet und erhielten nach jeder Screeningrunde einen Brief, der ihre Testergebnisse enthielt. Dieses Wissen über die Testergebnisse könnte dazu geführt haben, dass sich Teilnehmer außerhalb der Studie weiteren diagnostischen Untersuchungen unterzogen, um Gewissheit über ihren Gesundheitszustand zu erlangen. Dies bewirkte allerdings, dass Personen aus der Röntgen- oder LDCT-Gruppe CT-Untersuchungen erhielten und somit das Testergebnis überprüft wurde. Ausgewertet wurden die Teilnehmer aber nach der Intention-to-screen Methode, also zu der Gruppe zu der sie randomisiert wurden. Dadurch konnten die Screeningverfahren bessere Ergebnisse erzielen, obwohl dies durch andere Untersuchungen ermöglicht wurde. Zudem konnten die Radiologen klinische Befunde der Teilnehmer einsehen und ein bereits abgeklärter Befund, der als benigne bewertet wurde, wurde in den weiteren Screening-Runden nicht mehr als positiv deklariert. Damit konnten die Ärzte von den zusätzlichen diagnostischen Untersuchungen, die außerhalb der Studie durchgeführt wurden, profitieren. Verschiedene Wissenschaftler publizierten 2014 in JAMA einen Artikel, der die Überdiagnose Lungenkrebs im NLST einschätzte. Dabei wurde geschätzt dass 18,5% der detektierten Lungenkarzinome im LDCT eine Überdiagnose darstellten [59]. Dies sind jedoch lediglich Schätzungen, die erst durch eine längere Nachverfolgungszeit verifiziert werden könnten, denn dann müsste im Idealfall ohne Überdiagnose die Anzahl an Lungenkarzinomen, die mit dem Screening entdeckt werden, gleich der Anzahl an Lungenkarzinomen ohne Screening sein.

Ein weiteres Problem, dass bei dieser Studie auffällt ist, dass nur ein geringer Bruchteil der positiven Screeningbefunde durch eine Biopsie überprüft wurden. Es wurden meist weitere bildgebende Verfahren eingesetzt, um die

Wahrscheinlichkeit auf einen richtig positiven Befund zu erhöhen, bevor komplikationsreiche invasive Verfahren eingesetzt wurden. Dieses Vorgehen des NLST ist durchaus verständlich, wenn man berücksichtigt, dass mehr als jeder Dritte im Verlauf des Screenings mindestens ein positives Ergebnis erhielt und die Rate an falsch positiven Befunden beim LDCT bei 96,4% lag. Dadurch, dass nicht jeder, der einen positiven Befund beim Screening erhielt auch einer Biopsie unterzogen wurde, lässt wiederum Spielraum für falsch negative Befunde bei den nachfolgenden bildgebenden Verfahren, die dann möglicherweise erst Jahre später als Lungenkarzinom entdeckt wurden.

In vielen Screeningstudien, wie auch in dieser, wird oft als Endpunkt die „disease specific mortality“ gewählt. Doch ist die Wahl dieses Endpunktes überhaupt sinnvoll? Die Teilnehmer in dieser Studie befinden sich in einer Altersspanne, in der die häufigste Todesursache durch Herz-Kreislauferkrankungen verursacht wird (siehe dazu Abbildung 10 im Anhang). Die Entscheidung, ob ein Lungenkrebspatient an seiner Karzinomerkrankung oder an einer anderen Erkrankung gestorben war, ist in der Praxis oftmals schwierig eindeutig zu klären. Diese Zuordnung der Todesursache wurde im NLST durch ein Endpoint Verification Team vorgenommen, das aus fünf unabhängigen Ärzten bestand. Anhand des Totenscheins wurde eine Zuteilung auf Grundlage der Kriterien vorgenommen, die bereits bei Frage 9 dargestellt wurden. Dieses Verfahren ermöglicht dem EVT einen Handlungsspielraum und damit die Möglichkeit der Zuteilung zu Gunsten des Screenings.

Ein Ansatz, um die Zuordnung der Todesursache zu umgehen, könnte die Wahl eines anderen Endpunktes sein, und zwar die Bestimmung der Stadienverteilung der Karzinomerkrankung nach einer ausreichend langen Nachbeobachtungszeit. Bei einer effektiven Screeningmaßnahme kann davon ausgegangen werden, dass sich die Anzahl der Karzinome in einem hohen Stadium (T IV) im Vergleich zu einer Population ohne Screening verringern sollte. Dieser Endpunkt wäre eindeutig zu bestimmen und lässt wenig Platz für Manipulationen und Fehlentscheidungen.

Die Dauer der Nachbeobachtungszeit von durchschnittlich 6,5 Jahren ist äußerst knapp bemessen worden, zumal der Lead time bias auf 1,3 Jahre geschätzt wurde. Um Ergebnisse einer Studie sicher mit Daten zum 5-Jahres-Überleben und

mit Stadienverteilungen der Lungenkarzinome mit und ohne Screening vergleichen zu können, wäre eine Nachbeobachtungszeit von mindestens 10 Jahren sinnvoll.

In der LDCT-Gruppe sind circa 50% der entdeckten Karzinome im Stadium I. Dies könnte ein Hinweis darauf sein, dass es sich hierbei um einen Length time bias gehandelt haben könnte, denn langsam wachsende indolente Karzinome haben eine höhere Wahrscheinlichkeit im Stadium I diagnostiziert zu werden, als schnell wachsende aggressive Tumore. Diese langsam wachsenden Tumore weißen eine gute Prognose auf und können somit zu einem besseren Ergebnis beim Studienoutcome geführt haben.

Eine weitere markante Tatsache, die auffällt, bezieht sich auf die histologische Typenverteilung des Lungenkarzinoms im NLST. Das Plattenepithelkarzinom, das zu den NSCLC gehört, wurde nicht gesondert aufgeführt, obwohl bei diesem histologischen Typ ein Zusammenhang mit inhalativen Noxen gesichert ist. Der häufigste aufgetretene Typ war das Adenokarzinom mit 46,8% in der LDCT-Gruppe. In der Bevölkerung außerhalb der Studie nimmt das Adenokarzinom 35% unter den histologischen Typen ein. Damit wurden verhältnismäßig mehr Adenokarzinome, welche gehäuft bei Nichtrauchern und Frauen auftreten, innerhalb der Studie entdeckt. Diese Imbalance des Adenokarzinoms kann ein Hinweis auf eine Überdiagnose und möglicherweise auch auf einen Confounder sein, denn es besteht ein bewiesener direkter Zusammenhang für das Plattenepithelkarzinom und das Rauchen, wenn auch zugleich berücksichtigt werden muss, dass das Adenokarzinom auch bei Rauchern vorkommt und dadurch verursacht werden kann.

Diese Ergebnisse, die mithilfe der Fragensammlung entdeckt wurden, konnten nur unter Zuhilfenahme des Studienberichts (2011), des Studienprotokolls, des Papers „Study design und Overview“, des Artikels mit der Veröffentlichung der initialen Resultate nach dem T0 Screening (2013) und den jeweils dazugehörigen Appendices gefunden werden. Viele der Leser einer Studie beschäftigen sich aus Zeitgründen allein mit dem publizierten Studienbericht. Wird ausschließlich der Studienbericht vom NLST von 2011 für die Studienanalyse mit dem hier vorgestellten Fragebogen verwendet, dann sind folgende Fragen nicht zu beantworten: 1, 3.1, 3.2, 4.1, 8.2, und 9.2. Zudem kann Frage 2 und Frage 4.2 nur abgeschätzt werden. Somit besteht eine weitere Limitierung der Analyse einer

Studie mit der entwickelten Fragensammlung, wenn diese lediglich auf den Studienbericht angewendet wird, denn dadurch kann nur eine grobe Abschätzung der Wertigkeit einer Studie bezogen auf die Validität und klinische Relevanz getroffen werden.

Zusammenfassend muss die Existenz der diskutierten Bias angenommen oder zumindest vermutet werden. Der LDCT-Screeningtest weist unter alleiniger Berücksichtigung der berechneten Gütewerte eine interne Validität auf, wenn auch betont werden muss, dass sich der Test besser für den Ausschluss einer Erkrankung als für deren Nachweis eignet. Unter der Begutachtung aller aufgeführten Parameter kann die interne Validität des Tests nicht sicher angenommen werden. Die externe Validität des Screeningtests muss angezweifelt werden, da zum einen keine klaren Empfehlungen für die Bewertung eines Screeningbefundes existierten und zum anderen bei der Durchführung des Screeningprogramms Unterschiede zwischen den Studienzentren und zwischen Kontroll- und Experimentalgruppe bestand. Die klinische Relevanz lässt sich aus dem Vergleich vom Nutzen und Schaden für die Patienten und der Kosteneffektivität bestimmen. Die Sterblichkeit an Lungenkarzinom war in der LDCT-Gruppe um 20% geringer, was alleine durch die unterschiedlichen Ausgangsrisiken erklärbar ist. Drückt man den Unterschied in Absolutzahlen aus, so bedeutet das, dass sich die beiden Gruppen um drei pro 1000 Patienten unterschieden, die sich einem Screening mit LDCT oder mit Röntgen unterzogen haben.

Die Kosten für das Screening (wenn verschiedene Formen von Bias ausgeschlossen werden können) summieren sich auf 3074 US-Dollar pro Patient und bei der durchgeführten Diagnostik ohne Screening bei symptomatischem Lungenkrebs ergibt sich ein Wert von 1443 US-Dollar pro Patient. Aus der Differenz lassen sich die Mehrkosten für das Screening mit einem Wert von 1631 US-Dollar angeben. Dies ergibt eine ICER (Incremental Cost-Effectiveness Ratio) von 52000 (1631 x 0,0316) US-Dollar pro gewonnenem Lebensjahr und durch die Verrechnung mit 0,0201 qualitätskorrigierten Lebensjahren (QALY) ergibt sich ein Wert von 81000 US-Dollar pro QALY. Wie viel Geld letztendlich für das Screening ausgegeben werden will, hängt von Expertenmeinungen ab. Häufig wird in Europa eine Grenze bei 50000 Euro pro QALY gesetzt [19]. Diese berechneten Werte sind

abhängig von der Wahl der Hochrisikogruppe, von Therapiekosten, den Kosten für ein LDCT und ganz besonders von der Anzahl an Personen, die durch das Screening gerettet werden. Aus der Zusammenschau der diskutierten Bias in der Studie muss die Anzahl an Personen, die durch ein LDCT-Screening gerettet werden können, angezweifelt werden und somit wird die berechnete Kosteneffektivität irrelevant [77].

Letztendlich ist es fraglich, ob diese Studie eine klinische Relevanz besitzt. Da bereits in einigen Kliniken in Deutschland das Lungenkarzinom-Screening als Pilotprojekt eingeführt wurde, müssen die erzielten Ergebnisse den Nutzen und die Komplikationen für die Patienten und die Kosten für das Gesundheitswesen kritisch hinterfragt und geprüft werden, bevor eine bundesweite Entscheidung für die Implementierung getroffen werden kann.

5. Zusammenfassung

In dieser Arbeit wurde eine Sammlung von Fragen ausgearbeitet, mit welchen die Validität von Screeningstudien überprüft werden kann. Die Fragen mussten Bias (systematische Fehler) entdecken, die in Studien zur sekundären Prävention auftreten. Diese vorgeschlagene Sammlung von Fragen besteht aus 26 Fragen, wovon sich 16 auf den Nachweis von neun systematischen Fehlern und zehn auf die Bewertung des Verfahrens beziehen.

Die Fragensammlung wurde auf die im New England Journal of Medicine im Jahr 2011 publizierte Screeningstudie „Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening“ angewandt. In dieser Studie wird gezeigt, dass die Mortalität an Lungenkrebs durch das Screening mit Low-Dose Computertomographie gegenüber einem traditionellen Röntgen-Screening signifikant reduziert werden kann.

Die Validität dieser Studie wurde unter anderem durch folgende Faktoren eingeschränkt. Die Studienteilnehmer wurden von Anreizen (Werbung für die neue Screeningmethode) zur Teilnahme angeworben, was einen „healthy volunteer effect“ (gesündere Probanden) und einen „incentive bias“ (Patienten erhalten mehr Gesundheitsleistungen) begünstigt. Bereits zu Beginn der Studie waren die beiden Studienarme nicht vergleichbar, da in der Röntgen-Screening-Gruppe zum Zeitpunkt T0 56% mehr Patienten ein Lungenkarzinom im Stadium IV aufwiesen als in der Low-Dose Computertomographie-Gruppe (Selection Bias). Die Radiologen erhielten keine exakten Empfehlungen für die Beurteilung der radiologischen Befunde, womit sich die hohe Rate der falsch-positiven Befunde (96,4% in der Low-Dose Computertomographie-Gruppe und 94,5% in der Röntgen-Screening-Gruppe) erklären lässt. Die Ärzte bekamen eine spezielle Unterweisung und Zertifizierung für die Interpretation des Bildmaterials; zudem wurde der Zugriff auf klinische Patienteninformationen gewährt, sodass dadurch die Trefferquote für richtig-positive Befunde erhöht wurde (likelihood ratio bias). In den Zentren des American College of Radiology Imaging Network wurden weitere Untersuchungen (Blut-, Sputum- und Urinuntersuchungen) und eine häufigere

Befragungen der Patienten vorgenommen, als in den Lung Screening Study-Zentren. Dieser Unterschied in der Durchführung der Studie kann als „performance bias“ bezeichnet werden. Die Überdiagnosen von Lungenkarzinomen, welche durch das Low-Dose Computertomographie-Screening entdeckt werden, wird auf 18,5% geschätzt. Überdies fand weder eine Verblindung der Ärzte noch der Teilnehmer statt. Aus den Berechnungen zur Güte des Screeningtests lässt sich ableiten, dass sich dieser eher für den Ausschluss einer Krankheit als für deren Nachweis eignet (T0: Sensitivität: 93,8%, Spezifität: 73,4%).

Zusammenfassend ist sowohl die interne wie auch die externe Validität des Low-Dose Computertomographie-Screenings zur Reduktion der Sterblichkeit an Lungenkarzinom anzuzweifeln. Die klinische Relevanz, die sich aus der externen Validität des Tests, aus dem Nutzen für die Patienten und aus den Kosten für das Gesundheitswesen zusammensetzt, ist auf Grund dieser Studie nur mit Einschränkung zu bestätigen. Zukünftige Ergebnisse aus Kliniken, die dieses Screening anwenden, werden weitere Erkenntnisse liefern, um zu entscheiden, ob die Ergebnisse der vorliegenden Studie durch systematische Fehler verzerrt waren.

Obwohl ständig neue Screening- und Diagnostikstudien entwickelt werden, bleibt die Entwicklung von Methoden zur Überprüfung dieser Studien hinter der erstgenannten zurück.

Das medizinische Curriculum und die Standardlehrwerke der Inneren Medizin messen der Evaluation etablierter Screeningverfahren hinsichtlich des Nutzens und der Risiken keine große Bedeutung zu, sodass in diesem Bereich Nachholbedarf besteht.

6. Literaturverzeichnis

1. Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, Gareen IF, Gatsonis C, Marcus PM, Sicks JD: Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. N ENGL J MED 365: 395-409 (2011 a)

2. Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, Gareen IF, Gatsonis C, Marcus PM, Sicks JD: Supplementary Appendix. Supplement to: Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. N ENGL J MED 365: 395-409 (2011 b)

3. Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, Gareen IF, Gatsonis C, Marcus PM, Sicks JD: Protocol. Protocol for: Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. N ENGL J MED 365: 395-409 (2011 c)

4. Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, Gareen IF, Gatsonis C, Marcus PM, Sicks JD: Overview and Study Design. Radiology 258: 243-253 (2011 d)

5. Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, Gareen IF, Gatsonis C, Marcus PM, Sicks JD: Supplementary Appendix. Supplement to: Overview and Study Design. Radiology 258: 243-253 (2011 e)

6. American Cancer Society: http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-survival-by-stage (03.05.15)

7. Andreas S, Rittmeyer A, Hinterthaner M, Huber RM: Smoking Cessation in Lung Cancer-Achievable and Effective. Dtsch Arztebl Int 110: 719-724 (2013)

8. Approbationsordnung für Ärzte:

http://www.medizin.hhu.de/fileadmin/redaktion/Fakultaeten/Medizinische_Fakultaet/Studiendekanat/Dokumente/Rahmenbedingungen/AEAppO_zum__01_01_2014.pdf (03.05.15)

9. Bach PB, Mirkin JN, Oliver TK, Azzoli CG, Berry DA, Brawley OW, Byers T, Colditz GA, Gould MK, Jett JR, Sabichi AL, Smith-Bindman R, Wood DE, Qaseem A, Detterbeck FC: Benefits and Harms of CT Screening for Lung Cancer. JAMA 307: 2418-2429 (2012)

10. Berk RA: An introduction to sample selection bias in sociological data. American sociological review 48: 386-398 (1983)

11. Bhopal RS: über Variation. Role of errors, bias, and confounding. In: Bhopal RS, Bruce A, Usher J (Hrsg) Concepts of Epidemiology: An integrated introduction to the ideas, theories, principles and methods of epidemiology. Oxford University Press Oxford New York. 1. Auflage, S. 69-86 (2002)

12. Black WC, Gareen IF, Soneji SS, Sicks JD, Keeler EB, Aberle DR, Naeim A, Church TR, Silvestri GA, Gorelick J, Gatsonis C: Cost-Effectiveness of CT Screening in the National Lung Screening Trial. N ENGL J MED 371: 1793-1802 (2014)

13. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, De Vet HCW: Towards Complete and Accurate Reporting of Studies of Diagnostic Accuracy: The STARD Initiative. Clinical Chemistry 49: 1-6 (2003)

14. Bundesamt für Gesundheit: Evidence-based public health. http://www.henet.ch/ebph/09_bias/bias_091.php (03.05.15)

15. Canadian Task Force: The Periodic Health Examination. Canadian Medical Association Journal 121: 1193-1254 (1979)

16. Church TR, Black WC, Aberle DR, Berg CD, Clingan KL, Duan F, Fagerstrom RM, Gareen IF, Gierada DS, Jones GC, Mahon I, Marcus PM, Sicks JD, Jain A, Baum S: Results of Initial Low-Dose Computed Tomographic Screening for Lung Cancer. N ENGL J MED 368: 1980-1991 (2013)

17. Critical Appraisal Skills Programm: www.casp-uk.net (03.05.15)

18. Deutsches Ärzteblatt: Bundesrat stoppt Präventionsgesetz und Regelungen zur Korruptionsbekämpfung. Erschienen am 20.09.2013. http://www.aerzteblatt.de/nachrichten/55915/Bundesrat-stoppt-Praeventionsgesetz-und-Regelungen-zur-Korruptionsbekaempfung (03.05.15)

19. Deutsches Ärzteblatt: Lungenkrebs: Kosteneffektivität für CT-Früherkennung grenzwertig. Erschienen am 11.11.2014. http://www.aerzteblatt.de/nachrichten/60816/Lungenkrebs-Kosteneffektivitaet-fuer-CT-Frueherkennung-grenzwertig (03.05.15)

20. Dumville JC, Torgerson DJ, Hewitt CE: Reporting attrition in randomised controlled trials. BMJ 332: 969-971 (2006)

21. Etzioni R, Penson DF, Legler JM, di Tommaso D, Boer R, Gann PH, Feuer EJ: Overdiagnosis Due to Prostate-Specific Antigen Screening: Lessons From U.S. Prostate Cancer Incidence Trends. Journal of the National Cancer Institute 94: 981-990 (2002)

22. Fry WA, Menck HR, Winchester DP: The National Cancer Data Base Report on Lung Cancer. American Cancer Society 77: 1947-1955 (1996)

23. Gatsonis C: Do We Need a Checklist for Reporting the Results of Diagnostic Test Evaluations? The STARD Proposal. Acad Radiol 10: 599-600 Editorial (2003)

24. Gatsonis C, Paliwal P: Meta-Analysis of Diagnostic and Screening Test Accuracy Evaluations: Methodologic Primer. AJR 187: 271-281 (2006)

25. Gerok W, Huber C, Meinertz T, Zeidler H: über Grundlagen der Inneren Medizin. In: Gerok W, Huber C, Meinertz T, Zeidler H (Hrsg) Die Innere Medizin. Schattauer GmbH Stuttgart. 11.Auflage, S. 18 (2007)

26. Grill M, Hackenbroch V: Unsinn in bester Qualität. Der Spiegel Hamburg. Erschienen am 21.07.14, Nr. 30, S. 100-104 (2014)

27. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M: A framework for clinical evaluation of diagnostic technologies. CAN MED ASSOC J 134: 587-594 (1986)

28. Hahn JM: über Grundlagen der geriatrischen Diagnostik und Therapie. In: Bob A, Bob K (Hrsg) Duale Reihe: Innere Medizin. Thieme Stuttgart. 2.Auflage, S. 1427 (2009)

29. Hammerschmidt S, Wirtz H: Lungenkarzinom-aktuelle Diagnostik und Therapie. Dtsch Arztebl Int 106: 809-820 (2009)

30. Hansen WE: über die Bewertung von diagnostischen Untersuchungen. In: Berdel WE, Böhm M, Classen M, Diehl V, Kochsiek K, Schmiegel W (Hrsg) Innere Medzin. Urban & Fischer München/Jena. 5. Auflage, S. 7 (2004)

31. Harris RP, Sheridan SL, Lewis CL, Barclay C, Vu MB, Kistler CE, Golin CE, DeFrank JT, Brewer NT: The Harms of Screening. A Proposed Taxonomy and Application to Lung Cancer Screening. JAMA Internal Medicine 174: 281-285 (2014)

32. Herold G: über Bronchialkarzinom. In: Herold G (Hrsg) Innere Medzin. Dr. med. Gerd Herold Köln, S. 381-386 (2010)

33. Hewitt C, Hahn S, Torgerson DJ, Watson J, Bland M: Adequacy and reporting of allocation concealment: review of recent trials published in four general medical

journals. BMJ 330: 1057-1058 (2005)

34. Holland W: Screening: reasons to be cautious. Stringent criteria must be met before implementation. BMJ 306: 1222-1223 Editorial (1993)

35. Hollis S, Campbell F: What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ 319: 670-674 (1999)

36. Kubik AK, Parkin DM, Zatloukal P: Czech Study on Lung Cancer Screening. American Cancer Society 89: 2363-2368 (2000)

37. Lehnert H, Werdan K (Hrsg): Innere Medizin. Thieme Stuttgart/New York. 4. Auflage (2006)

38. van Leuween PJ, Kranse R, Hakulinen T, Roobol MJ, de Koning HJ, Bangma CH, Schröder FH: Disease-specific mortality may underestimate the total effect of prostate cancer screening. J Med Screen 17: 204-210 (2010)

39. Linnet K: A Review on the Methodology for Assessing Diagnostic Tests. CLIN. CHEM. 34: 1379-1386 (1988)

40. Linnet K, Bossuyt PM, Moons KG, Reitsma JB: Quantifying the Accuracy of a Diagnostic Test or Marker. CLIN. CHEM. 58: 1292-1301 (2012)

41. Löwe B: über diagnostische Kriterien. In: Bob A, Bob K (Hrsg) Duale Reihe: Innere Medizin. Thieme Stuttgart. 2.Auflage, S. 1391 (2009)

42. Luo S, Ainslie G, Giragosian L, Monterosso JR: Behavorial and Neural Evidence of Incentive Bias for Immediate Rewards Relative to Preference-Matched Delayed Rewards. The Journal of Neuroscience 29: 14820-14827 (2009)

43. Mahnken JD, Chan W, Freeman DH, Freeman Jr and JL: Reducing the effects of lead-time bias, length bias and over-detection in evaluating screening. Stat

Methods Med Res 17: 643-663 (2008)

44. Martin GJ: über Screening und Prävention von Krankheiten. In: Dietel M, Suttorp N, Zeitz M (Hrsg) Harrisons Innere Medizin. ABW-Wissenschaftsverlag Berlin. Bd 1, 17.Auflage, S. 30-33 (2008)

45. Melamed MR, Flehinger BJ, Zaman MB, Heelan RT, Perchick WA, Martini N: Screening for Early Lung Cancer. Results of the Memorial Sloan-Kettering Study in New York. CHEST 86: 44-53 (1984)

46. Michigan Center for Public Health Prepardness: http://practice.sph.umich.edu/micphp/epicentral/confounding.php (03.05.15)

47. Miller AB: Overdiagnosis of Breast Cancer. Int. J. Cancer 133: 2511 Editorial (2013)

48. Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG: CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 340: 1-28 (2010)

49. Morabia A, Zhang FF: History of medical screening: from concepts to action. Postgrad Med J 80: 463-469 (2004)

50. Moscucci O: The British Fight against Cancer: Publicity and Education, 1900-1948. Oxford University Press-Social History of Medicine 23: 356-373 (2009)

51. Nagel M: ICCT- The Internal Consistency of Clinical Trials. Med Dissertation. Universität Ulm, Klinische Ökonomik (2007)

52. Obuchowski NA, Graham RJ, Baker ME, Powell KA: Ten Criteria for Effective Screening: Their Application to Multislice CT Screening for Pulmonary and Colorectal Cancers. AJR 176: 1357-1362 (2001)

53. Ochodo EA, Bossuyt PM: Reporting the Accuracy of Diagnostic Tests: The STARD Initiative 10 Years On. Clinical Chemistry 59: 917-919 (2013)

54. ÖGD Baden-Württemberg: www.gesundheitsamt-bw.de/oegd/Gesundheitsthemen/Praevention (03.05.15)

55. Oken MM, Hocking WG, Kvale PA, Andriole GL, Buys SS, Church TR, Crawford ED, Fouad MN, Isaacs C, Reding DJ, Weissfeld JL, Yokochi LA, O`Brien B, Ragard LR, Rathmell JM, Riley TL, Wright P, Caparaso N, Hu P, Izmirlian G, Pinsky PF, Prorok PC, Kramer BS, Miller AB, Gohagan JK, Berg CD: Screening by Chest Radiograph and Lung Cancer Mortality. The Prostate, Lung, Colorectal, and Ovarian (PLCO) Randomized Trial. JAMA 306: 1865-1873 (2011)

56. Onkopedia Leitlinie für Lungenkarzinom, NSCLC. Deutsche Gesellschaft für Hämatologie und Medizinische Onkologie e.V. Berlin: https://dgho-onkopedia.de/de/onkopedia/leitlinien/lungenkarzinom-nicht-kleinzellig-nsclc/lungenkarzinom-nicht-kleinzellig-nsclc/index_html (03.05.15)

57. Otto SJ, Schröder FH, de Koning HJ: Low all-cause mortality in the volunteer-based Rotterdam section of the European randomised study of screening for prostate cancer: self-selection bias?. J Med Screen 11:89-92 (2004)

58. Patz EF, Caporaso NE, Dubinett SM, Massion PP, Hirsch FR, Minna JD, Gatsonis C, Duan F, Adams A, Apgar C, Medina RM, Aberle DR: National Lung Cancer Screening Trial American College of Radiology Imaging Network Specimen Biorepository Originating from the Contemporary Screening for the Detection of Lung Cancer Trial (NLST, ACRIN 6654). Journal of Thoracic Oncology 5: 1502-1506 (2010)

59. Patz EF, Pinsky P; Gatsonis C, Sicks JD, Kramer BS, Tammemägi MC, Chiles C, Black WC, Aberle DR : Overdiagnosis in Low-Dose Computed Tomography Screening for Lung Cancer. JAMA Intern Med 174: 269-274 (2014)

60. Peto R, Darby S, Deo H, Silcocks P, Whitley E, Doll E: Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies. BMJ 321: 323-329 (2000)

61. Pinsky PF, Miller A, Kramer BS, Church T, Reding D, Prorok P, Gelmann E, Schoen RE, Buys S, Hayes RB, Berg CD: Evidence of a Healthy Volunteer Effect in the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial. Am J Epidemiol 165: 874-881 (2007)

62. Piper W (Hrsg): Innere Medizin. Springer Medizin Berlin/Heidelberg/New York. 1. Auflage (2007)

63. Plint AC, Moher D, Morrison A, Schulz K, Altman DG, Hill C, Gaboury I: Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. Medical Journal of Australia 185: 263-267 (2006)

64. Porzsolt F, Du Prel J-B, Muttray A: über Inhalte der klinischen Ökonomik. In: Porzsolt F (Hrsg.) Grundlagen der Klinischen Ökonomik. PVS Berlin. 1. Auflage, S. 35-61 (2011)

65. Porzsolt F: Begriffshygiene-der erste Schritt zu mehr Effizienz in der Gesundheitsversorgung. DER GELBE DIENST. Gesundheits-und Sozialpolitik – Nachrichten, Analysen, Hintergrund 13: 9-11 (2014)

66. Porzsolt F, Braubach P, Flurschütz PI, Göller A, Sailer MB, Weiss M, Wyer P: Medical Students Can Help Avoid the Expert Bias in Medicine. Scientific Research. Creative Education 3: 1115-1121 (2012)

67. Porzsolt F, Bausch J, Geipel G, Huppertz E, Mühlbacher A, Otto T, Radic D, Schmidt P, Ravens-Sieberer U, Zimmermann TM, Clouth J: über die angemessene Evidenz für Therapieentscheidungen: eine Diskussion des Methodenpluralismus in klinischen Studien. In: Rychlik R, Erdmann E, Rebscher H, Selbmann HK, Straub

C, Ulrich V, Wille E (Hrsg) Gesundheitsökonomie & Qualitätsmanagement. Gesundh ökon Qual manag 18: 31-39, Thieme Stuttgart/ New York (2013)

68. PRISMA: www.prisma-statement.org (03.05.15)

69. Puliti D, Duffy SW, Miccinesi G, de Koning H, Lynge E, Zappa M, Paci E: Overdiagnosis in mammographic screening for breast cancer in Europe: a literature review. J Med Screen 19: 42-56 (2012)

70. Raffle A, Gray JAM: über was Screening ist und was es nicht ist. In: Raffle A, Gray JAM, Piribauer F, Gartlehner G, Mad P, Waechter F (Hrsg) Screening Durchführung und Nutzen von Vorsorgeuntersuchungen. Hans Huber, Hogrefe AG Bern. 1. Auflage, S. 53-60 (2009)

71. Ransohoff DF, Feinstein AR: Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 299: 926-930 (1978)

72. Reid MC, Lachs MS, Feinstein AR: Use of methodological standards in diagnostic test research. Getting better but still not good. JAMA 274: 645-651 (1995)

73. Renz-Polster H, Krautzig S (Hrsg): Basislehrbuch Innere Medizin. Elsevier GmbH, Urban & Fischer München/Jena. 4. Auflage (2008)

74. Riegelman RK: Conduct of Tests. In: Riegelmann RK (Hrsg) Studying a Study &Testing a Test. How to read the medical evidence. Lippincott Williams & Wilkins Philadelphia. 5. Auflage, S. 146-147 (2005)

75. Robert Koch Institut: Zentrum für Krebsregisterdaten. www.krebsdaten.de/Krebs/DE/Content/Krebsarten/Lungenkrebs/lungenkrebs_node.html (03.05.15)

76. Robra B-P: über klinische Epidemiologie. In: Classen M, Diehl V, Kochsiek K

(Hrsg) Innere Medizin. Urban & Schwarzenberg München/Wien/Baltimore. 3. Auflage, S. 3-5 (1994)

77. Rösch M-C, Porzsolt F: Validität wichtiger als Kosten. Kommentar zu Deutsches Ärzteblatt: Lungenkrebs: Kosteneffektivität für CT-Früherkennung grenzwertig. Erschienen am 11.11.2014. http://www.aerzteblatt.de/nachrichten/60816/Lungenkrebs-Kosteneffektivitaet-fuer-CT-Frueherkennung-grenzwertig (03.05.15)

78. Rose G, Barker DJ: Screening. BMJ 2: 1417-1418 (1978)

79. Sauerland S, Neugebauer E: Evidenzbasierte Chirurgie. In: Krukemeyer MG, Spiegel H-U (Hrsg) Chirurgische Forschung. Thieme Stuttgart. 1. Auflage, S. 230-233 (2006)

80. Schmidt M: über Tumoren der Bronchien und der Lunge. In: Bob A, Bob K (Hrsg) Duale Reihe: Innere Medizin. Thieme Stuttgart. 2.Auflage, S. 407-416 (2009)

81. Schmidt RL, Factor RE: Understanding Sources of Bias in Diagnostic Accuracy Studies. Arch Pathol Lab Med 137: 558-565 (2013)

82. Schmiegel W: über Prävention. In: Berdel WE, Böhm M, Classen M, Diehl V, Kochsiek K, Schmiegel W (Hrsg) Innere Medzin. Urban & Fischer München/Jena. 5. Auflage, S. 201-203 (2004)

83. Schulz KF, Altman DG, Moher D: CONSORT 2010: Aktualisierte Leitlinie für Berichte randomisierter Studien im Parallelgruppen-Design. Dtsch Med Wochenschr 136: e20-e23 (2011)

84. Selman TJ, Morris RK, Zamora J, Khan KS: The quality of reporting of primary test accuracy studies in obstetrics and gynaecology: application of the STARD criteria. Selman et al. BMC Women`s Health 11: 1-7 (2011)

85. Statistica.
http://de.statista.com/statistik/daten/studie/182666/umfrage/computertomographen-anzahl-in-europa/ (03.05.15)

86. Statistica.
http://de.statista.com/statistik/daten/studie/166516/umfrage/krankenhaeuser-mit-medizinisch-technischen-grossgeraeten/ (03.05.15)

87. Statistisches Bundesamt. Todesursachen in Deutschland 2012.
https://www.destatis.de/DE/Publikationen/Thematisch/Gesundheit/Todesursachen/Todesursachen2120400127004.pdf?__blob=publicationFile (03.05.15)

88. Stobbe H: über diagnostische Grundlagen. In: Stobbe H, Baumann G (Hrsg) Innere Medizin. Ullstein Mobsy GmbH & Co.KG Berlin/Wiesbaden. 7. Auflage, S. 6-7 (1996)

89. Strauss GM, Gleason RE, Sugarbaker DJ: Screening for Cancer Re-examined. A Reinterpretation of the Mayo Lung Project Randomized Trial on Lung Cancer Screening. CHEST 103: 337S-341S (1993)

90. Takiguchi Y, Sekine I, Iwasawa S: Overdiagnosis in Lung Cancer Screening with Low-Dose Computed Tomography. Journal of Thoracic Oncology 8: e101-e102, Letter to the Editor (2013)

91. Thornton H, Edwards A, Baum M: Women need better information about routine mammography. BMJ 327: 101-103 (2003)

92. Tockman MS: Survival and mortality from lung cancer in a screened population: The Johns Hopkins Study. CHEST 89: 324S-325S (1986)

93. Tripepi G, Jager KJ, Dekker FW, Zoccali C: Selection Bias and Information Bias in Clinical Research. Nephron Clin Pract 115: c94-c99 (2010)

94. Turner L, Shamseer L, Altman DG, Schulz KF, Moher D: Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review. Systematic Reviews 1: 60 (2012)

95. Waber RL, Shiv B, Carmon Z, Ariely D: Commercial Features of Placebo and Therapeutic Efficacy. JAMA 299: 1016-1017 (2008)

96. Weiß C, Rzany B: über Epidemiologische Studien. In: Weiß C (Hrsg) Basiswissen Medizinische Statistik. Springer Heidelberg. 5.Auflage, S. 256-262 (2010 a)

97. Weiß C, Rzany B: über Risikostudien. In: Weiß C (Hrsg) Basiswissen Medizinische Statistik. Springer Heidelberg. 5. Auflage, S. 265-283 (2010 b)

98. Weiß C, Rzany B: über Studien zu Diagnostik und Prävention. In: Weiß C (Hrsg) Basiswissen Medizinische Statistik. Springer Heidelberg. 5. Auflage, S. 287-302 (2010 c)

99. Welch HG: über you may have a „cancer scare“ and face an endless cycle of testing. In: Welch HG (Hrsg) Should i be tested for cancer? Maybe not and here`s why. University of California Press. 1. Auflage, S. 33-50 (2004)

100. Welch HG, Woloshin S, Schwartz LM, Gordis L, Gotzsche PC, Harris R, Kramer BS, Ransohoff DF: Overstating the Evidence for Lung Cancer Screening. The International Early Lung Cnacer Action Program (I-ELCAP) Study. Arch Intern Med. 167: 2289-2295 (2007)

101. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J: The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC 3: 25 (2003)

102. Wild F: Nutzen der Prävention im Gesundheitswesen – ein Literaturüberblick. WIP Diskussionspapier 6/07 Köln, S. 1-11 (2007). http://www.wip-pkv.de/uploads/tx_nppresscenter/Nutzen_der_Praevention_im_Gesundheitswesen.pdf (03.05.15)

103. Wilson JM, Jungner G: über Definitions. In: Wilson JM, Jungner G (Hrsg) Principles and Practice of Screening for Disease. Public Health Papers 34. World Health Organization France/Switzerland/Geneva, S. 11-13 (1968)

104. Windeler J: Externe Validität. Z. Evid. Fortbild. Qual. Gesundh. wesen (ZEFQ) 102: 253-260 (2008)

105. Yen AM-F, Duffy SW, Chen TH-H, Chen L-S, Chiu SY-H, Fann JC-Y, Wu WY-Y, Su C-W, Smith RA, Tabar L: Long-Term Incidence of Breast Cancer by Trial Arm in One County of Swedish Two-County Trial of Mammographic Screening. Cancer 118: 5728-5732 (2012)

7. Anhang

Tabelle 2: CONSORT Checkliste [83]

In der folgenden Tabelle ist die veröffentlichte CONSORT Checkliste abgebildet, die entwickelt wurde um Autoren einer Studie eine Leitlinie für die korrekte Berichterstattung zu geben.
(CONSORT: Consolidated Statement of Reporting Trials)

Section/Topic	Item No	Checklist item	Reported on page No
Title and abstract			
	1a	Identification as a randomised trial in the title	
	1b	Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts)	
Introduction			
Background and objectives	2a	Scientific background and explanation of rationale	
	2b	Specific objectives or hypotheses	
Methods			
Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio	
	3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons	
Participants	4a	Eligibility criteria for participants	
	4b	Settings and locations where the data were collected	
Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered	
Outcomes	6a	Completely defined pre-specified primary	

		and secondary outcome measures, including how and when they were assessed	
	6b	Any changes to trial outcomes after the trial commenced, with reasons	
Sample size	7a	How sample size was determined	
	7b	When applicable, explanation of any interim analyses and stopping guidelines	
Randomisation:			
Sequence generation	8a	Method used to generate the random allocation sequence	
	8b	Type of randomisation; details of any restriction (such as blocking and block size)	
Allocation concealment mechanism	9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned	
Implementation	10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions	
Blinding	11a	If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how	
	11b	If relevant, description of the similarity of interventions	
Statistical methods	12a	Statistical methods used to compare groups for primary and secondary outcomes	
	12b	Methods for additional analyses, such as subgroup analyses and adjusted analyses	
Results			
Participant flow (a diagram is strongly recommended)	13a	For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome	

	13b	For each group, losses and exclusions after randomisation, together with reasons	
Recruitment	14a	Dates defining the periods of recruitment and follow-up	
	14b	Why the trial ended or was stopped	
Baseline data	15	A table showing baseline demographic and clinical characteristics for each group	
Numbers analysed	16	For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups	
Outcomes and estimation	17a	For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)	
	17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended	
Ancillary analyses	18	Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory	
Harms	19	All important harms or unintended effects in each group (for specific guidance see CONSORT for harms)	
Discussion			
Limitations	20	Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses	
Generalisability	21	Generalisability (external validity, applicability) of the trial findings	
Interpretation	22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence	
Other information			
Registration	23	Registration number and name of trial registry	
Protocol	24	Where the full trial protocol can be accessed, if available	

Funding	25	Sources of funding and other support (such as supply of drugs), role of funders	

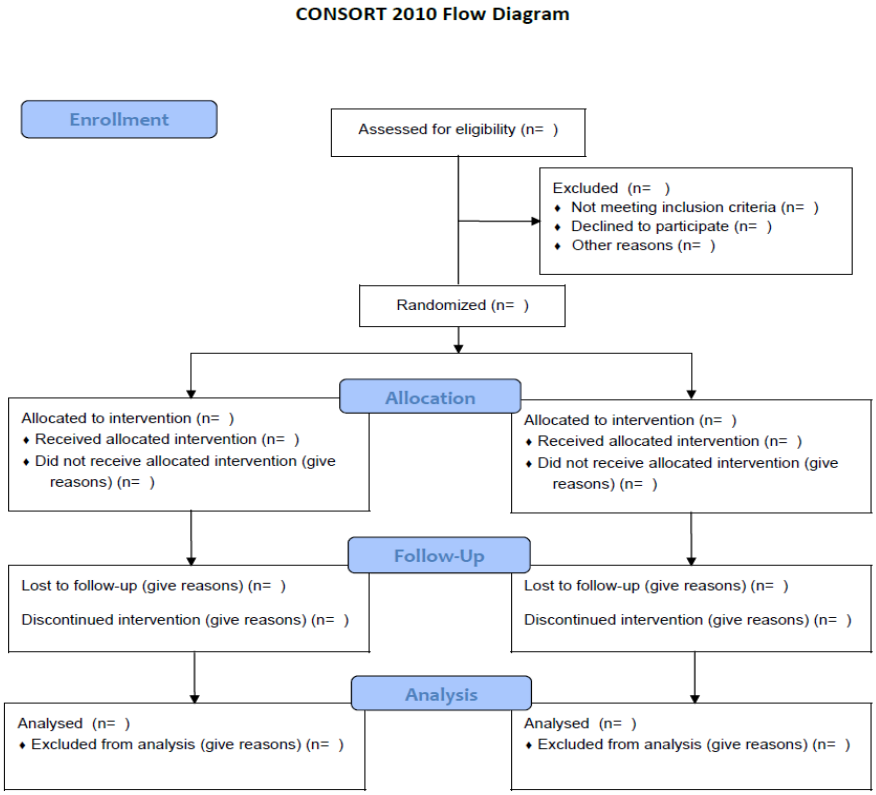

Abbildung 1: CONSORT Flussdiagramm [83]

Die obige Abbildung soll den Verlauf der Studienteilnehmer von der Aufnahme in die Studie bis zur finalen Auswertung ihrer Studienergebnisse zeigen. Dieses Flussdiagramm soll für die notwendige Transparenz in der Studienberichterstattung sorgen.

(CONSORT: Consolidated Statement of Reporting Trials)

Tabelle 3: STARD Checkliste [13]

Die folgende Checkliste wurde entwickelt, um mit deren Hilfe die Qualität eines Screening- oder Diagnostiktests zu messen.

(STARD: Standards for Reporting of Diagnostic Accuracy)

Section and Topic	Item #		On page #
TITLE/ABSTRACT/ KEYWORDS	1	Identify the article as a study of diagnostic accuracy (recommend MeSH heading 'sensitivity and specificity').	
INTRODUCTION	2	State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.	
METHODS			
Participants	3	The study population: The inclusion and exclusion criteria, setting and locations where data were collected.	
	4	Participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?	
	5	Participant sampling: Was the study population a consecutive series of participants defined by the selection criteria in item 3 and 4? If not, specify how participants were further selected.	
	6	Data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?	
Test methods	7	The reference standard and its rationale.	
	8	Technical specifications of material and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard.	
	9	Definition of and rationale for the units, cut-offs and/or categories of the results of the index tests and the reference standard.	
	10	The number, training and expertise of the persons	

		executing and reading the index tests and the reference standard.	
	11	Whether or not the readers of the index tests and reference standard were blind (masked) to the results of the other test and describe any other clinical information available to the readers.	
Statistical methods	12	Methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals).	
	13	Methods for calculating test reproducibility, if done.	
RESULTS			
Participants	14	When study was performed, including beginning and end dates of recruitment.	
	15	Clinical and demographic characteristics of the study population (at least information on age, gender, spectrum of presenting symptoms).	
	16	The number of participants satisfying the criteria for inclusion who did or did not undergo the index tests and/or the reference standard; describe why participants failed to undergo either test (a flow diagram is strongly recommended).	
Test results	17	Time-interval between the index tests and the reference standard, and any treatment administered in between.	
	18	Distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.	
	19	A cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard.	
	20	Any adverse events from performing the index tests or the reference standard.	
Estimates	21	Estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals).	
	22	How indeterminate results, missing data and outliers of the index tests were handled.	
	23	Estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done.	
	24	Estimates of test reproducibility, if done.	

DISCUSSION	25	Discuss the clinical applicability of the study findings.	

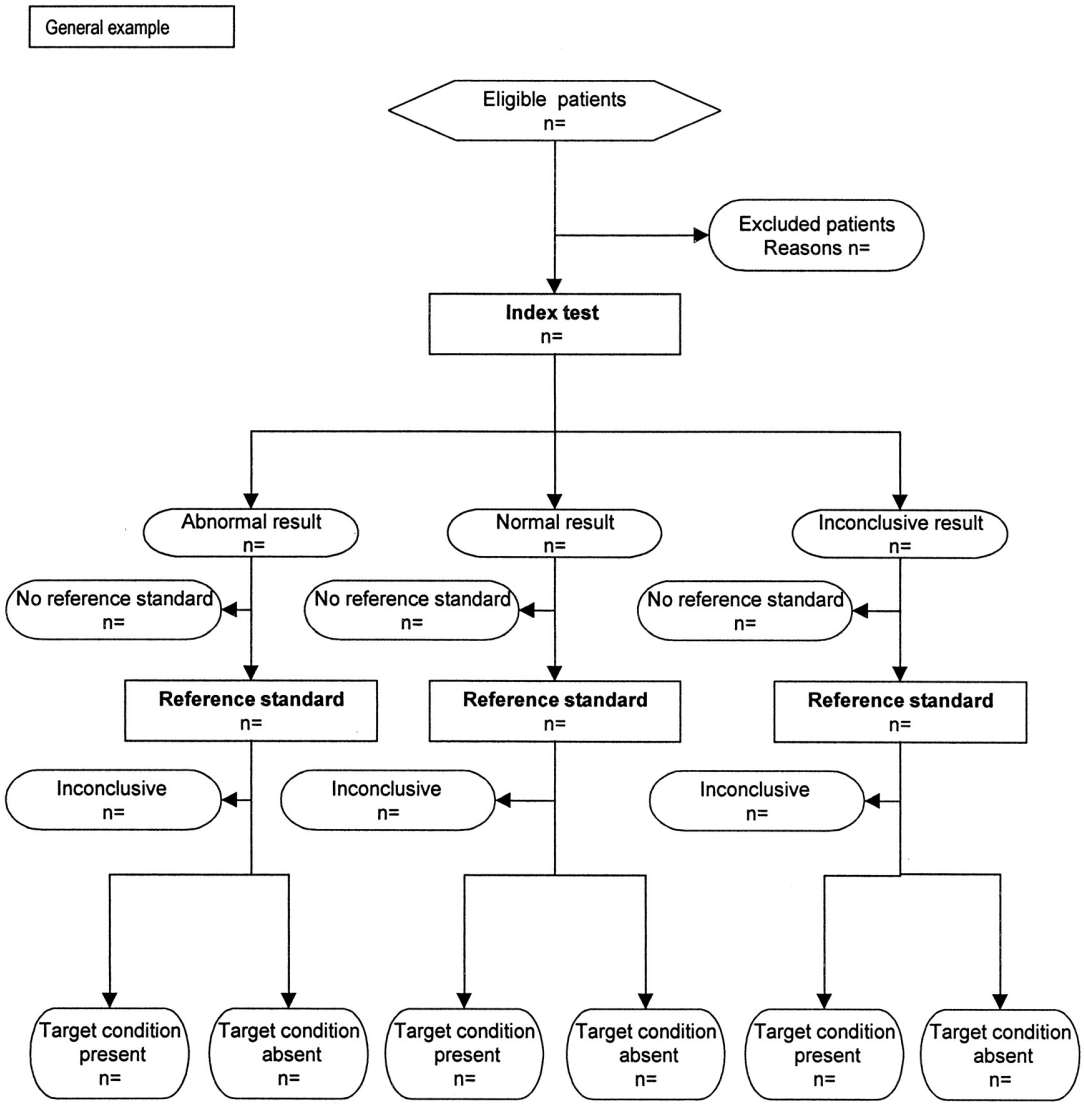

Abbildung 2: STARD Flussdiagramm [13]

Die obige Abbildung zeigt ein Flussdiagramm, welches für Diagnostikstudien entwickelt wurde. Dieses Flussdiagramm gibt Informationen zur Patientenrekrutierung, Ausführung des Tests und über die Anzahl an Patienten, welche den Index- und Referenztest durchlaufen haben, an.

(STARD: Standards for Reporting of Diagnostic Accuracy. Indextest: Der Test, dessen Nutzen und Überlegenheit in einer Studie gezeigt werden soll. Referenztest: Der aktuell beste verfügbare Test zum Nachweis des gesuchten Zusammenhangs.)

(**Lizenz:** Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Bossuyt PM, Reitsma JB, Bouter LM, de Vet HC: „Reproducibility of the STARD checklist: an instrument to assess the quality of reporting of diagnostic accuracy studies". DOI: 10.1186/1471-2288-6-12. CC-BY 2.0. http://creativecommons.org/licenses/by/2.0)

Tabelle 4: Checkliste von QUADAS [101]

Die folgende Checkliste besteht aus 14 Themen, die zu 14 Fragen formuliert wurden. Dieser Fragebogen stellt ein Instrument dar mit welchem Diagnostikstudien bewertet werden können und dient als eine Methode zur Qualitätsprüfung.

(QUADAS: Quality Assessment of Diagnostic Accuracy Studies)

(**Lizenz**: Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PN, Kleijnen J: „Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies“. DOI: 10.1186/1471-2288-6-9. CC-BY 2.0. http://creativecommons.org/licenses/by/2.0)

Item		Yes	No	Unclear
1.	Was the spectrum of patients representative of the patients who will receive the test in practice?	( )	( )	( )
2.	Were selection criteria clearly described?	( )	( )	( )
3.	Is the reference standard likely to correctly classify the target condition?	( )	( )	( )
4.	Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?	( )	( )	( )
5.	Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis?	( )	( )	( )
6.	Did patients receive the same reference standard regardless of the index test result?	( )	( )	( )
7.	Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?	( )	( )	( )
8.	Was the execution of the index test described in sufficient detail to permit replication of the test?	( )	( )	( )
9.	Was the execution of the reference standard described in sufficient detail to permit its replication?	( )	( )	( )
10.	Were the index test results interpreted without knowledge of the results of the reference standard?	( )	( )	( )
11.	Were the reference standard results interpreted without knowledge of the results of the index test?	( )	( )	( )
12.	Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?	( )	( )	( )
13.	Were uninterpretable/ intermediate test results reported?	( )	( )	( )
14.	Were withdrawals from the study explained?	( )	( )	( )

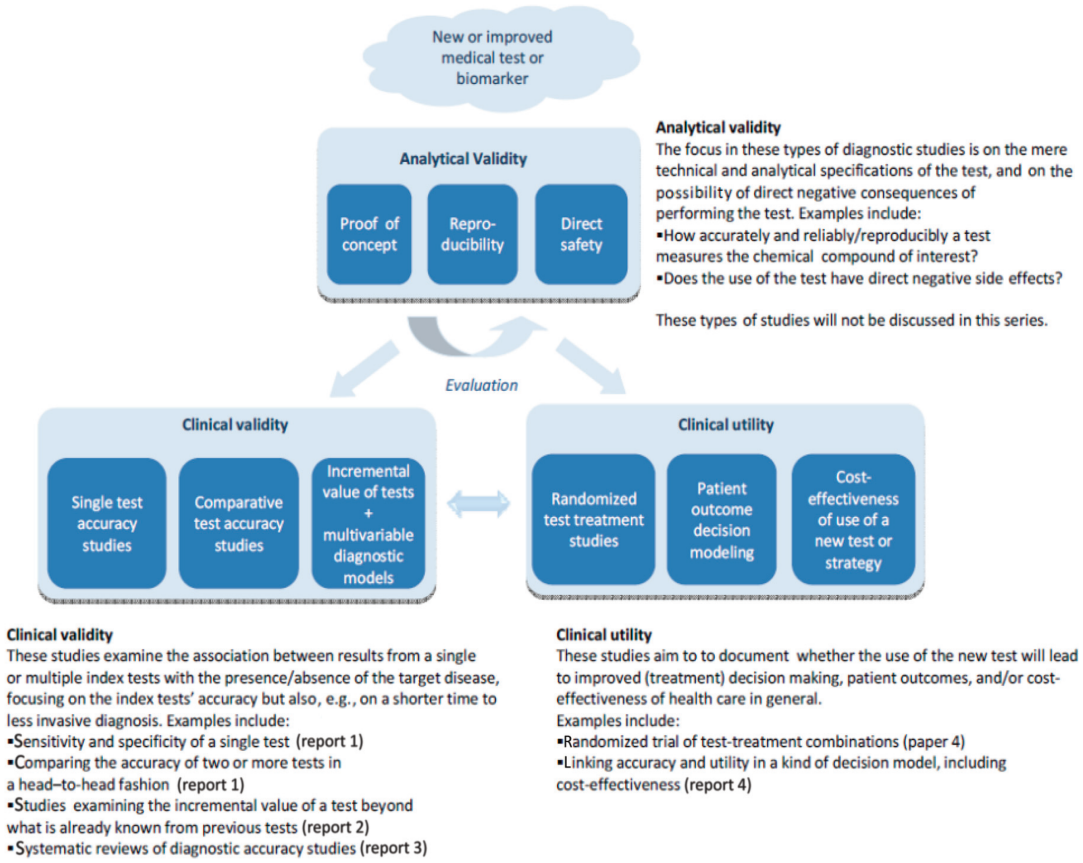

Abbildung 3: Übersicht über die Hauptunterteilung für die Bewertung von diagnostischen Tests [40]

Dieses Schaubild fasst drei Hautkomponenten zusammen, die bedeutend sind um die Überlegenheit eines neuen diagnostischen Tests zu erfassen. Dabei wird besonders darauf geachtet ob ein Test eine analytische Validität, einen klinischen Nutzen und eine klinische Validität besitzt.

Tabelle 6: Auswertung der Screeningstudie des National Lung Screening Trial (NLST) mit dem USP (Usability of scientific publication)-Fragebogen [66]

Die Validität der Studie (Frage 10) ergibt sich aus der Summation der Fragen 1-9. Dabei wird jeder Frage, bei zutreffender Antwort ein Punkt zugeteilt. Die erste Frage enthält zwei Kriterien, weshalb jedes Kriterium einen halben Punkt bekommt. Unter 4 Punkten ist die Studie nicht valide. Zwischen 4 und 6,5 Punkten ist die Studie eingeschränkt valide. Ab 7 Punkten wird die Studie als valide bezeichnet. Die klinische Relevanz ergibt sich aus der externen Validität (Übertragbarkeit der Studie), Nutzen für den Patienten und aus dem Kosten-Nutzen-Prinzip für die gesetzlichen Krankenkassen.

Frage	**Fragen zur Prüfung der Kriterien**	zutreffend	Nicht zutreffend
1	Sind die Ziele und Endpunkte der Studie benannt?	x	
2	Wurde das ideale Studiendesign gewählt?	x	
3	Sind die Risiken der Studienpopulation ähnlich?	x*	
4	War die Allokation zu den Studiengruppen geheim?		x
5	Wurde die Untersuchung bei Personen mit definierten Risiken oder bei Personen mit bestimmten Symptomen durchgeführt?	x	
6	War die Nachbeobachtungszeit ausreichend lange, um aussagekräftige Endpunkte messen zu können?	x (einge-schränkt)	
7	Wurden alle Patienten in die Analyse einbezogen?	x	
8	Ist die angewandte Statistik angemessen?	x	
9	Wurden Interessenkonflikte ausgeschlossen?		x
10	Halten Sie die Aussage der Studie für valide?	x	
11	Ist der Effekt der Studie klinisch relevant?	eingeschränkt relevant	

* Die Risiken in den beiden Studienarmen sind laut der Tabelle 1 auf Seite 399 im Studienbericht des NLST [1] ähnlich.

Zu Frage 10: Die Fragen 1-9 summieren sich auf 7 Punkte. Damit ist die Studie, laut dem USP-Fragebogen eindeutig valide. Zählt man die diskussionswürdigen Fragen als nicht zutreffend, ergibt sich eine Summe von 5 Punkten und damit zeigt sich eine eingeschränkte Validität der Studie.

Tabelle 7: Erweiterter USP (Usability of scientific publication)-Fragebogen

Der ursprüngliche USP-Fragebogen wurde um Items erweitert, die für die Bewertung eines Screeningverfahrens entscheidend sind. Diese Items wurden im Anschluss an die Frage elf angehängt.

Die kursiv unterlegten und mit einem *- markierten Fragen wurden zu den indirekten Bias für die Fragensammlung für die Evaluierung von Screeningstudien zusammengefasst (Tabelle 10).

Frage	**Fragen zur Prüfung der Kriterien**	zutreffend	Nicht zutreffend
1	Sind die Ziele und Endpunkte der Studie benannt?		
2	*Wurde das ideale Studiendesign gewählt? **		
3	Sind die Risiken der Studienpopulation ähnlich?		
4	*War die Allokation zu den Studiengruppen geheim? **		
5	Wurde die Untersuchung bei Personen mit definierten Risiken oder bei Personen mit bestimmten Symptomen durchgeführt?		
6	War die Nachbeobachtungszeit ausreichend lange, um aussagekräftige Endpunkte messen zu können?		
7	Wurden alle Patienten in die Analyse einbezogen?		
8	*Ist die angewandte Statistik angemessen? **		
9	*Wurden Interessenkonflikte ausgeschlossen? **		
10	Halten Sie die Aussage der Studie für valide?		
11	Ist der Effekt der Studie klinisch relevant?		
	Repräsentatives Spektrum von Personen?		
	*Verblindung aller Beteiligten in der Studie (z.B. Ärzte, Teilnehmer)? **		
	Referenzstandard akzeptabel?		
	Zeitraum zwischen den Tests akzeptabel		
	Wurde der Referenzstandard unabhängig vom Testergebnis des Indextests angewendet?		
	Wurde der zu untersuchende Test in einer zweiten, unabhängigen Gruppe von Patienten validiert?		
	Differentielle Verifikation vermieden?		

	Referenzstandard Ergebnisse geblindet?		
	Indextest Ergebnisse geblindet?		
	Nicht interpretierbare Ergebnisse berichtet?		
	Widerrufung der Teilnahme erklärt?		
	Prospektive Datenerhebung?		
	Cut-off Punkte für Indextest definiert?		
	Reproduzierbarkeit des untersuchten Tests gegeben?		
	Screening-Gruppe: am Screening teilgenommen n=?		
	Screening-Gruppe: am Screening nicht teilgenommen n=?		
	Kontroll-Gruppe: am Screening teilgenommen n=?		
	Kontroll-Gruppe: am Screening nicht teilgenommen n=?		
	Zeit-Intervall Indextest bis Therapie?		
	Zeit-Intervall Referenztest bis Therapie?		
	Schweregrade der Erkrankung in den Studienarmen *-Einheitlich?* *-Mehrere Subgruppen?* *-Daten nach Subgruppen aufgeführt? **		
	Sensitivität des Screeningtests/Referenztests?		
	Spezifität des Screeningtests/Referenztests?		
	Positive und negative Likelihood Ratio des Screeningtests/Referenztest?s		
	*Unerwünschte Wirkungen des Screeningtests? **		
	Existiert eine Vierfeldertafel für die produzierten Ergebnisse?		
	Statistik für mehrere Patienten-Subgruppen angegeben?		
	*Ist bei einem Vergleich zweier Indextests, der zu vergleichende Indextest akzeptabel?**		
	Wird der Indextest das weitere Management verändern?		
	Liefert der Indextest wichtige Informationen, wie zum Beispiel. eine Prognose für die Erkrankung?		
	Prä-Test Wahrscheinlichkeit der Erkrankung?		

	Kosten des Indextests angegeben?		
	Verfügbarkeit des Screeningtests? *		
	Prävalenz der Erkrankung? Prävalenz unterschiedlicher Schweregrade?		
	Prognose der Erkrankung? Prognose unterschiedlicher Schweregrade?		
	Verfügbarkeit effektiver Therapien? *		
	Schadenspotential effektiver Therapien?		
	Lebensqualität bei effektiver Therapie?		

Tabelle 12: Vierfeldertafeln für die T1-Screeningrunde (Näherung)

Mit dieser Vierfeldertafel lassen sich Aussagen zur Testgenauigkeit, Sensitivität, Spezifität, positiv und negativ prädiktivem Wert für die T1-Screeningrunde dieser Studie treffen.

Die Zahlenwerte wurden denselben Tabellen und Abbildungen entnommen wie bei Tabelle 11 im Ergebnisteil 3.3 beschrieben.

(CT: Computertomograph. CT positiv / Röntgen positiv: Der Untersuchungsbefund deutet auf die gesuchte Erkrankung hin. CT negativ / Röntgen negativ: Der Untersuchungsbefund deutet nicht auf die gesuchte Erkrankung hin. Krank: Personen mit der gesuchten Erkrankung, wobei Σ TP+FN die Gesamtzahl der Kranken angibt. Gesund: Personen ohne die gesuchte Erkrankung, wobei Σ FP+TN die Gesamtzahl der Gesunden angibt. TP („True Positive“): Anzahl der Patienten, die nach einem positiven Befund auch tatsächlich an der gesuchten Erkrankung leiden. FP („False Positive“): Anzahl der Patienten, die nach einem positiven Befund nicht an der gesuchten Erkrankung leiden. FN („False Negative“): Anzahl der Personen, die nach einem negativen Befund trotzdem an der gesuchten Erkrankung leiden. TN („True Negative“): Anzahl der Personen, die nach einem negativen Befund nicht an der gesuchten Erkrankung leiden. PPW: positiv prädiktiver Wert. NPW: negativ prädiktiver Wert. PLHR: positive Likelihood Ratio. NLHR: negative Likelihood Ratio.)

CT-Gruppe

	Krank	***Gesund***	
CT positiv	168 (TP)	6733 (FP)	6901
CT negativ	18 (FN)	17796 (TN)	17814
	186	24529	24715

Röntgen-Gruppe

	Krank	***Gesund***	
Röntgen positiv	65 (TP)	1417 (FP)	1482
Röntgen negativ	68 (FN)	22539 (TN)	22607
	133	23956	24089

Sensitivität (CT) = 168 / 186 = 90,3 %

Spezifität (CT) = 17796 / 24529 = 72,6 %

PPW (CT) = 168 / 6901 = 2,4 %

NPW (CT) = 17796 / 17814 = 99,9 %

PLHR (CT) = 0,903 / (1 – 0,726) = 3,3

NLHR (CT) = (1 – 0,903) / 0,726 = 0,13

Testgenauigkeit (CT) = (168 + 17796) / 24715 = 0,73

Sensitivität (Röntgen) = 65 / 133 = 48,9 %

Spezifität (Röntgen) = 22539 / 23956 = 94,1 %

PPW (Röntgen) = 65 / 1482 = 4,4 %

NPW (Röntgen) = 22539 / 22607 = 99,7 %

PLHR (Röntgen) = 0,489 / (1 – 0,941) = 8,3

NLHR (Röntgen) = (1 – 0,489) / 0,941 = 0,54

Testgenauigkeit (Röntgen) = (65 + 22539) / 24089 = 0,94

Tabelle 13: Vierfeldertafeln für die T2-Screeningrunde (Näherung)

Mit dieser Vierfeldertafel lassen sich Aussagen zur Testgenauigkeit, Sensitivität, Spezifität, positiv und negativ prädiktivem Wert für die T2-Screeningrunde dieser Studie treffen.

Die Zahlenwerte wurden denselben Tabellen und Abbildungen entnommen wie bei Tabelle 11 im Ergebnisteil 3.3 beschrieben.

(CT: Computertomograph. CT positiv / Röntgen positiv: Der Untersuchungsbefund deutet auf die gesuchte Erkrankung hin. CT negativ / Röntgen negativ: Der Untersuchungsbefund deutet nicht auf die gesuchte Erkrankung hin. Krank: Personen mit der gesuchten Erkrankung, wobei Σ TP+FN die Gesamtzahl der Kranken angibt. Gesund: Personen ohne die gesuchte Erkrankung, wobei Σ FP+TN die Gesamtzahl der Gesunden angibt. TP („True Positive“): Anzahl der Patienten, die nach einem positiven Befund auch tatsächlich an der gesuchten Erkrankung leiden. FP („False Positive“): Anzahl der Patienten, die nach einem positiven Befund nicht an der gesuchten Erkrankung leiden. FN („False Negative“): Anzahl der Personen, die nach einem negativen Befund trotzdem an der gesuchten Erkrankung leiden. TN („True Negative“): Anzahl der Personen, die nach einem negativen Befund nicht an der gesuchten Erkrankung leiden. PPW: positiv prädiktiver Wert. NPW: negativ prädiktiver Wert. PLHR: positive Likelihood Ratio. NLHR: negative Likelihood Ratio.)

Siehe Tabelle 11 und Tabelle 12.

CT-Gruppe

	Krank	***Gesund***	
CT positiv	211 (TP)	3843 (FP)	4054
CT negativ	26 (FN)	20022 (TN)	20048
	237	23865	24102

Röntgen-Gruppe

	Krank	***Gesund***	
Röntgen positiv	78 (TP)	1096 (FP)	1174
Röntgen negativ	66 (FN)	22106 (TN)	22172
	144	23202	23346

Sensitivität (CT) = 211 / 237 = 89,0 %

Spezifität (CT) = 20022 / 23865 = 83,9 %

PPW (CT) = 211 / 4054= 5,2 %

NPW (CT) = 20022 / 20048 = 99,9 %

PLHR (CT) = 0,89 / (1 – 0,839) = 5,5

NLHR (CT) = (1 – 0,89) / 0,839 = 0,13

Testgenauigkeit (CT) = (211 + 20022) / 24102 = 0,84

Sensitivität (Röntgen) = 78 / 144 = 54,2 %

Spezifität (Röntgen) = 22106 / 23202 = 95,3 %

PPW (Röntgen) = 78 / 1174 = 6,6 %

NPW (Röntgen) = 22106 / 22172 = 99,7 %

PLHR (Röntgen) = 0,542 / (1 – 0,953) = 11,5

NLHR (Röntgen) = (1 – 0,542) / 0,953 = 0,48

Testgenauigkeit (Röntgen) = (78 + 22106) / 23346 = 0,95

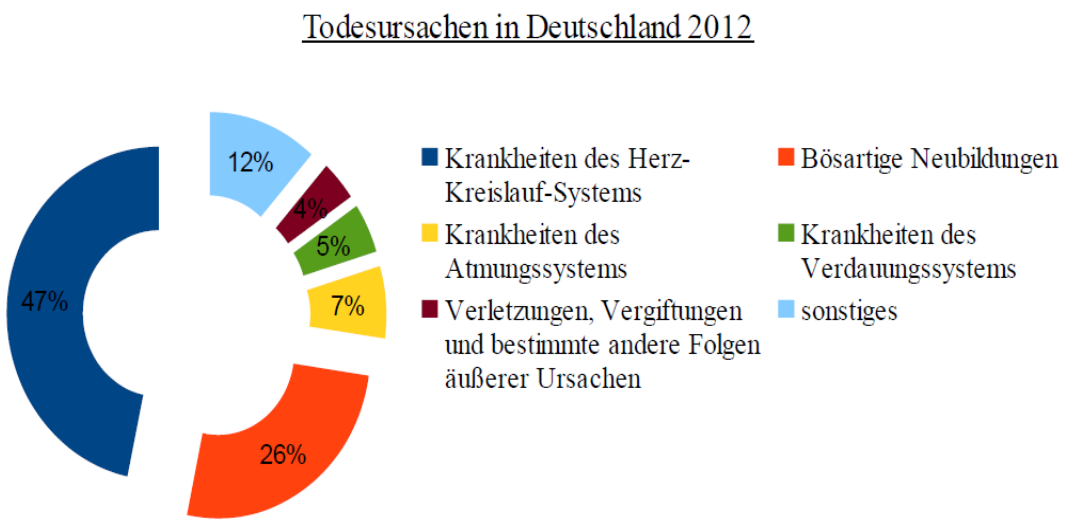

Abbildung 10: Todesursachen in Deutschland 2012

(eigene Darstellung, Quelle: Statistisches Bundesamt [87])

Diese Abbildung als Tortendiagramm dargestellt, zeigt das Verhältnis von bestimmten Todesursachen zur Gesamtzahl der Verstorbenen im Jahr 2012. Die verschiedenen Farben erlauben eine Zuordnung zu den einzelnen Todesursachen und sind rechts neben der Abbildung dargestellt.

Danksagung

Ich bedanke mich recht herzlich für die Unterstützung, die ich im Zuge der Dissertation von den Kolleginnen und Kollegen des Instituts erfahren habe.

Mein größter Dank gebührt für seine große Geduld und steten Hilfestellung Herrn Professor Dr. med. Franz Porzsolt, der sich stets Zeit für meine Fragen und Probleme nahm. Ohne seine Unterstützung und Motivation wäre diese Arbeit nicht möglich gewesen.

Ebenso möchte ich mich bei meinen Kollegen der Screening-Gruppe, Frau Isik und Herrn Seuffert, für die anregenden Diskussionen bedanken.

Des Weiteren möchte ich mich ganz besonders bei meiner Schwester für die große Unterstützung vor allem in der Endphase meiner Arbeit bedanken.

Zu guter Letzt möchte ich mich bei meinen Eltern und meinem Freund für die fortwährende Unterstützung während meines gesamten Studiums recht herzlich bedanken!

Lebenslauf

Curriculum vitae

Persönliche Daten:

Name:	Marie-Christin Rösch
Geburtsdaten:	25.02.1988 in Ulm

Schulbildung:

Sept. 1998 – Jun. 2007	Bertha-von Suttner Gymnasium Pfuhl
Jun. 2007	**Abitur**, Leistungsfächer Mathematik und Physik

Studium

Okt. 2007	Beginn Medizinstudium an der Universität Ulm
Aug. 2009	**Physikum**
Aug. 2012	Beginn des Praktischen Jahres am Klinikum Heidenheim
Okt. 2013	**2. Staatsexamen**
Jan 2014 – Okt. 2014	Assistenzärztin in der Anästhesiologie an der Klinik am Eichert in Göppingen
Jan. 2015	Assistenzärztin in der Anästhesiologie am Universitätsklinikum Ulm