DER EORTC QLQ-F17 ALS NEUER KERNFRAGEBOGEN ZUR ERFASSUNG DER FUNKTIONALEN LEBENSQUALITÄT ONKOLOGISCHER PATIENTEN

Untersuchung der Äquivalenz und psychometrischen Eigenschaften
im Rahmen einer randomisierten, multinationalen cross-over Studie

Dissertation
zur Erlangung des Doktorgrades
der Humanwissenschaften
(Dr. sc. hum.)

der
Fakultät für Medizin
der Universität Regensburg

vorgelegt von
Florian Zeman
aus
Weiden in der Oberpfalz

im Jahr
2025

DER EORTC QLQ-F17 ALS NEUER KERNFRAGEBOGEN ZUR ERFASSUNG DER FUNKTIONALEN LEBENSQUALITÄT ONKOLOGISCHER PATIENTEN

Untersuchung der Äquivalenz und psychometrischen Eigenschaften
im Rahmen einer randomisierten, multinationalen cross-over Studie

Dissertation
zur Erlangung des Doktorgrades
der Humanwissenschaften
(Dr. sc. hum.)

der
Fakultät für Medizin
der Universität Regensburg

vorgelegt von
Florian Zeman
aus
Weiden in der Oberpfalz

im Jahr
2025

Dekan:	Prof. Dr. Dirk Hellwig
Betreuer:	Prof. Dr. Michael Koller

Inhaltsverzeichnis

Zusammenfassung 6
Summary 8
1 Einleitung 10
1.1 Lebensqualität 11
1.1.1 Definition von Lebensqualität 11
1.1.2 Erfassung der Lebensqualität 12
1.2 EORTC 13
1.3 Fragebögen der EORTC 14
1.3.1 EORTC QLQ-C30 14
1.3.2 Module 14
1.3.3 Item Library 15
1.3.4 EORTC QLQ-F17 16
1.4 Äquivalenzstudien in der klinischen Forschung 19
1.5 Ziele der Studie 21
2 Material und Methoden 22
2.1 Studienplanung 22
2.1.1 Systematische Literaturrecherche 22
2.1.2 Konzept zur Datenerhebung 22
2.1.3 Entwicklung der Patientenbefragung 24
2.1.4 Vorstudie und Finalisierung der Patientenbefragung 25
2.1.5 Statistischer Analyseplan 25
2.2 Studiendurchführung 26
2.2.1 Studiendesign 26
2.2.2 Studienaufbau 26
2.2.3 Datenerhebung 28
2.3 Beschreibung der Messinstrumente 29
2.3.1 EORTC QLQ-C30 (Version 3.0) 29
2.3.2 EORTC QLQ-F17 30
2.3.3 Zwischenfragen 31
2.4 Methodische Definitionen 31
2.4.1 Äquivalenzgrenzen 31

2.4.2 Analysepopulation und Datenqualität ... 32
2.4.2.1 Qualitätsindikator: Konsistenz der ersten 7 Fragen ... 33
2.4.2.2 Qualitätsindikator: Plausibilität der Antwortzeiten ... 34
2.5 Fallzahlberechnung ... 35
2.6 Statistische Methoden ... 37
2.6.1 Deskriptive Statistiken ... 37
2.6.2 Berechnung der Skalenscores des QLQ-C30 und des QLQ-F17 ... 38
2.6.3 Überprüfung der Äquivalenz ... 39
2.6.3.1 Zwischen-Gruppenvergleiche ... 39
2.6.3.2 Innerhalb-Gruppenvergleiche ... 49
2.6.4 Subgruppenanalysen ... 53
2.6.5 Psychometrische Eigenschaften des QLQ-F17 ... 54
3 Ergebnisse ... 56
3.1 Definition der Analysepopulation ... 56
3.2 Beschreibung der Analysepopulation ... 63
3.3 Äquivalenz - Zwischen-Gruppenvergleiche ... 67
3.3.1 Differenzielle Item-Funktion ... 67
3.3.2 Multiple lineare Regressionsanalysen ... 76
3.4 Äquivalenz – Innerhalb-Gruppenvergleiche ... 77
3.4.1 Lineare gemischte Modelle ... 77
3.4.2 Unterstützende Analysen ... 79
3.5 Subgruppenanalysen ... 81
3.5.1 Symptomlast ... 81
3.5.2 Geschlecht ... 83
3.6 Psychometrische Eigenschaften des QLQ-F17 ... 85
3.7 Repräsentativität der Analysepopulation ... 89
4 Diskussion ... 92
4.1 Bewertung der Äquivalenz ... 93
4.2 Psychometrische Eigenschaften ... 97
4.3 Repräsentativität der Ergebnisse ... 98
4.4 Limitationen ... 99
4.4.1 Datenerhebung ... 99
4.4.2 Cross-over Design und Wash-out Phase ... 100

4.5 Perspektiven 101
5 Schlussfolgerung 102
6 Literaturverzeichnis 103
7 Anhang 109
A) EORTC QLQ-C30 Fragebogen in der Deutschen Originalform 109
B) Patientenbefragung 111
Tabellenverzeichnis 125
Abbildungsverzeichnis 127
Abkürzungsverzeichnis 128
Danksagung 130
Selbstständigkeitserklärung 131

ZUSAMMENFASSUNG

Die Erhebung der gesundheitsbezogenen Lebensqualität hat in der onkologischen Forschung eine hohe Relevanz. Sie umfasst sowohl das Handlungsvermögen der Patienten, also funktionale Aspekte wie beispielsweise Bewältigung des Alltags oder soziales Miteinander, als auch das subjektive Wohlbefinden, das sich unter anderem in körperlichen oder seelischen Symptomen äußern kann. Um diese Aspekte zuverlässig zu erfassen, kommen standardisierte Fragebögen zum Einsatz, von denen der EORTC QLQ-C30 zu den international etablierten Instrumenten zählt. Dieser enthält neben den funktionalen Skalen jedoch auch eine feste Auswahl an Symptomskalen, die nicht immer ideal auf den jeweiligen klinischen Kontext oder die spezifische Behandlungssituation zugeschnitten sind. Im Zuge der Forderung sowohl von behördlicher Seite wie der FDA als auch von Patientenvertretungen nach kürzeren und flexibleren Fragebögen zur Entlastung der Patienten wurde der QLQ-F17 als kompakterer Kernfragebogen konzipiert, der sich ausschließlich auf die funktionalen Dimensionen konzentriert.

Obwohl der QLQ-F17 und der QLQ-C30 dieselben funktionalen Dimensionen mit identischen Fragen erfassen, unterscheiden sie sich in ihrer Struktur. So wurden im QLQ-F17 die Symptomskalen entfernt, die im QLQ-C30 zwischen den funktionalen Skalen platziert sind. Dadurch verändert sich der inhaltliche Kontext der funktionalen Fragen im QLQ-F17, was potenzielle Auswirkungen auf das Antwortverhalten haben kann. Primäres Ziel der vorliegenden Arbeit war es somit, die Äquivalenz des neu entwickelten Kernfragebogens QLQ-F17 mit dem etablierten QLQ-C30 in Bezug auf die Erfassung funktionaler Lebensqualität und des globalen Gesundheitszustands bei onkologischen Patienten zu untersuchen. Zusätzlich wurden der direkte Einfluss der Symptomskalen auf das Antwortverhalten sowie die psychometrischen Eigenschaften des QLQ-F17 untersucht.

Im Rahmen einer randomisierten, multinationalen cross-over Studie wurden für die Untersuchung der Äquivalenz über 2.500 Patienten eingeschlossen. Die Äquivalenzprüfung erfolgte auf Skalen-Ebene mittels linearer Regressionsanalysen unter Verwendung vordefinierter Äquivalenzgrenzen sowie auf Item-Ebene mithilfe von DIF-Analysen. Die Ergebnisse der Analysen zeigten auf beiden Ebenen eine durchgehend konsistente Äquivalenz zwischen den beiden Fragebögen. Jedoch

zeigten sich leichte Unterschiede im Antwortverhalten bei einzelnen Items, insbesondere bei solchen, die im QLQ-C30 direkt auf die Symptomskalen folgen. Diese zwar sehr kleinen und klinisch nicht relevanten Unterschiede deuten dennoch auf potenzielle Kontexteffekte, etwa in Form von Ankereffekten, hin und entsprechen theoretischen Modellen zum Antwortverhalten in Fragebögen. Dies unterstreicht die Notwendigkeit, auch bei der Veränderung standardisierter Instrumente derartige Validierungsanalysen durchzuführen. Ergänzend zeigte die psychometrische Bewertung des QLQ-F17 eine gute Reliabilität und Validität im Vergleich zum QLQ-C30.

Insgesamt bestätigen die Ergebnisse die Äquivalenz der funktionalen Skalen zwischen dem QLQ-F17 und dem QLQ-C30 und sprechen dafür, den QLQ-F17 als kürzeres, fokussiertes Instrument zur Erhebung funktionaler Lebensqualität in zukünftigen onkologischen Studien einzusetzen. Als Kernfragebogen lässt sich der QLQ-F17 flexibel um symptombezogene Items aus der EORTC Item Library ergänzen, die an krankheitsspezifische Beschwerden und erwartbare Nebenwirkungen angepasst werden können. Der QLQ-F17 stellt damit ein verlässliches und zeiteffizientes Instrument zur Erhebung patientenberichteter Funktionsfähigkeit in der onkologischen Forschung und Versorgung dar.

SUMMARY

The assessment of health-related quality of life plays a vital role in oncological research. It encompasses both patients' functional abilities such as managing daily life or engaging in social interactions and their subjective well-being, which can be reflected in physical or psychological symptoms. To reliably capture these aspects, standardized questionnaires are used, among which the EORTC QLQ-C30 is one of the internationally established instruments. In addition to functional scales, however, the QLQ-C30 includes a fixed set of symptom scales, which are not always ideally suited to the specific clinical context or treatment situation. In response to demands from both regulatory bodies such as the FDA and patient advocacy groups for shorter and more flexible questionnaires to reduce patient burden, the QLQ-F17 was developed as a more compact core questionnaire focused exclusively on functional dimensions.

Although the QLQ-F17 and the QLQ-C30 measure the same functional dimensions using identical items, they differ in structure. In the QLQ-F17, the symptom scales that are placed between the functional scales in the QLQ-C30 have been removed. As a result, the contextual positioning of the functional items changes in the QLQ-F17, which may influence response behavior. The primary aim of this study was to examine the equivalence of the newly developed core questionnaire QLQ-F17 with the established QLQ-C30 in assessing functional quality of life and global health status in cancer patients. In addition, the direct influence of the symptom scales on response behavior as well as the psychometric properties of the QLQ-F17 were investigated.

In a randomized, multinational cross-over study, over 2,500 patients were included to evaluate equivalence. Equivalence was tested at the scale level using linear regression analyses with predefined equivalence margins, and at the item level using DIF analyses. The results showed consistent equivalence between the two questionnaires at both levels. However, slight differences in response behavior were observed for individual items, particularly those that follow directly after the symptom scales in the QLQ-C30. Although these differences were very small and not clinically relevant, they suggest potential context effects, such as anchoring effects, and are in line with theoretical models of questionnaire response behavior. This underlines the importance of conducting such validation analyses when modifying standardized instruments. In

addition, the psychometric evaluation of the QLQ-F17 demonstrated good reliability and validity compared to the QLQ-C30.

Overall, the findings confirm the equivalence of the functional scales between the QLQ-F17 and the QLQ-C30 and support the use of the QLQ-F17 as a shorter, more focused instrument for assessing functional quality of life in future oncological studies. As a core questionnaire, the QLQ-F17 can be flexibly extended with symptom-related items from the EORTC Item Library, which can be tailored to disease-specific complaints and anticipated side effects. The QLQ-F17 thus represents a reliable and time-efficient instrument for assessing patient-reported functioning in oncological research and care.

Vorbemerkung: Teile dieser Arbeit wurden bereits vorab in einem internationalen renommierten Fachjournal publiziert (1).

1 EINLEITUNG

Die gesundheitsbezogene Lebensqualität hat sich in den vergangenen Jahren zu einem wesentlichen Aspekt bei der Bewertung und Auswahl onkologischer Therapien entwickelt. Bereits in den 1970er Jahren begann sich die Erkenntnis durchzusetzen, dass das Überleben allein kein ausreichendes Maß für den Erfolg einer Krebstherapie ist. Mit der Entwicklung moderner Therapien, die zwar die Überlebenszeiten verlängern, aber häufig auch erhebliche Nebenwirkungen mit sich bringen, rückte die Lebensqualität (Quality of Life, QOL) zunehmend in den Fokus der Forschung (2).

Die Lebensqualität umfasst dabei nicht nur das körperliche Wohlbefinden, sondern geht über klassische medizinische Parameter hinaus und berücksichtigt das subjektive Erleben der Patienten hinsichtlich physischer, psychischer und sozialer Belastungen im Rahmen ihrer Erkrankung und Behandlung. Angesichts der oft belastenden Nebenwirkungen onkologischer Therapien sowie der mit der Diagnose verbundenen psychosozialen Herausforderungen kommt der Lebensqualität sowohl in klinischen Studien als auch in der individuellen Therapieplanung eine immer größere Bedeutung zu.

Gerade in fortgeschrittenen Stadien von Krebserkrankungen, bei denen eine Heilung nicht immer möglich ist, gewinnt die Lebensqualität als Therapieziel zunehmend an Gewicht. Doch auch bei potenziell heilbaren Tumorerkrankungen ist es wichtig, die Lebensqualität in die Entscheidungsfindung einzubeziehen, um unnötige Belastungen zu vermeiden und die Lebenszufriedenheit der Patienten zu fördern. Studien zeigen, dass die gesundheitsbezogene Lebensqualität nicht nur ein wichtiger Zielparameter ist, sondern darüber hinaus als unabhängiger prognostischer Faktor für das Überleben fungieren kann (3).

Durch die systematische Erhebung von Lebensqualitätsdaten bereits vor Beginn der Therapie lassen sich individuelle Problembereiche frühzeitig identifizieren. Dies ermöglicht es, unterstützende Maßnahmen, psychosoziale Interventionen oder begleitende Therapien gezielt einzusetzen, um die Belastungen während der Behandlung zu verringern. Auf diese Weise kann ein auf die Bedürfnisse und das

Befinden der einzelnen Patientin bzw. des einzelnen Patienten zugeschnittener Behandlungsplan entwickelt werden, der sowohl medizinische Wirksamkeit als auch persönliche Lebensqualität in den Mittelpunkt stellt.

1.1 LEBENSQUALITÄT

1.1.1 DEFINITION VON LEBENSQUALITÄT

Die Auffassung von Gesundheit hat sich in den letzten Jahrzehnten zunehmend erweitert. Die Weltgesundheitsorganisation WHO hat bereits 1946 in ihrer Verfassung festgelegt, dass unter Gesundheit nicht lediglich das Fehlen von Krankheit oder Beschwerden, sondern ein Zustand umfassenden körperlichen, seelischen und sozialen Wohlbefindens zu verstehen sei (4). Diese ganzheitliche Perspektive verdeutlicht, dass zur Bewertung von Gesundheit auch subjektive Empfindungen und soziale Lebensumstände einbezogen werden müssen, ein Gedanke, der eng mit dem Konzept der Lebensqualität verknüpft ist.

Bereits in den 1980er Jahren wurde versucht, die Lebensqualität begrifflich zu fassen. Der britische Onkologe Calman schlug vor, dass Lebensqualität durch die Diskrepanz zwischen dem, was ein Mensch erwartet, und dem, was er tatsächlich erlebt, bestimmt wird (5). Daraus ergibt sich, dass die Wahrnehmung der eigenen Lebensqualität stark individuell geprägt und dynamisch ist. Sie hängt von persönlichen Werten, Erfahrungen und Lebenszielen ab und kann sich im Verlauf einer Erkrankung oder Behandlung verändern.

Auch die WHO betont in ihrer Definition von Lebensqualität diesen subjektiven Charakter. Sie beschreibt sie als subjektive Wahrnehmung der eigenen Lebenssituation im Kontext kultureller und gesellschaftlicher Wertsysteme sowie im Hinblick auf persönliche Ziele, Erwartungen und Prioritäten (6). Dieser Ansatz betont die individuelle Bewertung des eigenen Lebens und berücksichtigt die persönlichen Rahmenbedingungen und Wertvorstellungen.

In der Medizin hat sich aus dieser allgemeinen Definition das Konzept der „gesundheitsbezogenen Lebensqualität“ (Health-Related Quality of Life, HRQOL) entwickelt. Es umfasst die durch Krankheit oder Therapie beeinflussten Aspekte des alltäglichen Lebens und integriert körperliche, psychische und soziale Dimensionen.

Entscheidend ist hierbei, dass die Bewertung dieser Bereiche aus der subjektiven Sicht der Patienten erfolgt (7). Die HRQOL stellt damit eine wichtige Ergänzung zu objektiv-medizinischen Kennzahlen dar und erlaubt eine umfassendere Bewertung des Behandlungserfolgs.

1.1.2 Erfassung der Lebensqualität

Die systematische Erfassung der Lebensqualität in der klinischen Forschung und Versorgung stellt eine methodische Herausforderung dar. Ein zentrales Problem liegt darin, dass Patienten sehr unterschiedliche Vorstellungen von Gesundheit und Lebensqualität mitbringen. Die individuelle Bewertung hängt stark von persönlichen Erwartungen, bisherigen Erfahrungen und der Fähigkeit zur Anpassung an gesundheitliche Einschränkungen ab. So kann ein Patient mit vergleichsweise milden Symptomen seine Lebensqualität als stark beeinträchtigt empfinden, während eine schwer erkrankte Person, die gelernt hat, mit ihrer Situation umzugehen, ihre Lebensqualität positiver einschätzt. Solche interindividuellen Unterschiede führen dazu, dass bei der Erhebung der Lebensqualität eines relativ homogenen Patientenkollektivs oft hohe Varianzen auftreten.

Ein weiteres zentrales Merkmal der gesundheitsbezogenen Lebensqualität (HRQOL) ist ihr subjektiver Charakter. Die Einschätzung erfolgt durch den Patienten selbst und kann daher nicht von außen objektiv bestimmt werden (8). Diese Form der Selbstbewertung wird als Patient-Reported Outcome (PRO) bezeichnet und bildet eine unverzichtbare Informationsquelle für eine patientenzentrierte Versorgung (9).

Dabei ist zu berücksichtigen, dass die persönliche Bewertung der Lebensqualität einem ständigen Wandel unterliegen kann. Insbesondere bei chronischen oder fortschreitenden Erkrankungen wie Krebserkrankungen verändert sich oft der individuelle Bezugsrahmen, nach dem Lebensqualität beurteilt wird. Auch kurzfristige Faktoren wie die aktuelle Stimmung oder emotionale Verfassung können die Antworten in einer Lebensqualitätsbefragung beeinflussen (10).

Angesichts dieser Komplexität erfordert die Messung der Lebensqualität einen multidimensionalen Ansatz. Um ein möglichst vollständiges Bild zu erhalten, sollten verschiedene Lebensbereiche berücksichtigt wie körperliche Funktionsfähigkeit, psychisches Wohlbefinden sowie soziale Teilhabe berücksichtigt werden (11). Nur

durch eine solche ganzheitliche Betrachtung lassen sich die Auswirkungen einer Erkrankung oder Therapie auf das Leben der Betroffenen realistisch erfassen und in die klinische Entscheidungsfindung einbeziehen.

1.2 EORTC

Die Europäische Organisation für Forschung und Behandlung von Krebs (EORTC) ist eine 1962 gegründete internationale, gemeinnützige Organisation mit Sitz in Brüssel, Belgien. Ihr Hauptziel ist es, die Behandlungsergebnisse und die Lebensqualität von Krebspatienten weltweit zu verbessern. Die EORTC vereint Wissenschaftler und Kliniker aus verschiedenen Disziplinen und Ländern, um innovative Studien und Projekte im Bereich der Onkologie durchzuführen. Sie fungiert als Schnittstelle zwischen Grundlagenforschung und klinischer Anwendung, indem sie Ergebnisse aus Laborstudien direkt in klinische Praktiken integriert. Durch ihre Arbeit trägt die EORTC nicht nur dazu bei, neue Behandlungsansätze zu entwickeln, sondern auch dazu, bestehende Therapien zu optimieren und deren Nebenwirkungen zu minimieren.

Die EORTC erkannte frühzeitig die Bedeutung standardisierter Instrumente zur Messung der Lebensqualität. So wurde im Jahr 1980 in Reaktion auf den Bedarf an einer einheitlichen Strategie für die Lebensqualitätsforschung innerhalb der EORTC die Quality of Life Group (QLG) gegründet. Ihre initiale Aufgabe bestand darin, die EORTC-Zentrale sowie die verschiedenen Kooperationsgruppen bei der Planung, Durchführung und Auswertung von Studien zur Lebensqualität im Rahmen ausgewählter Phase-III-Studien zu beraten. Dies war ein entscheidender Schritt, da es bis dato keine etablierten, universellen Instrumente gab, die die spezifischen Herausforderungen und Bedürfnisse von Krebspatienten angemessen abbilden konnten. Die QLG setzte sich von Beginn an aus einem interdisziplinären Team aus Onkologen, Radiotherapeuten, Chirurgen, Psychiater, Spezialisten für Palliativmedizin, Psychologen, Sozialarbeiter und Methodenexperten zusammen. Heute sind in der QLG Vertreter aus 43 europäischen Ländern sowie aus Australien, Kanada und den Vereinigten Staaten organisiert. Eine der aktuellen Kernaufgaben der QLG ist die Entwicklung und Anpassung von Fragebögen zur Erhebung der gesundheitsbezogenen Lebensqualität von Patienten zur Verwendung in onkologischen klinischen Studien und in der klinischen Praxis.

1.3 FRAGEBÖGEN DER EORTC

1.3.1 EORTC QLQ-C30

Der EORTC QLQ-C30 (Quality of Life Questionnaire-Core 30) wurde erstmals in den 1980er Jahren konzipiert und 1993 in seiner finalen Version veröffentlicht (12). Er gilt heute als einer der Goldstandards in der Lebensqualitätsforschung in der Onkologie. Der Fragebogen wurde speziell für Krebspatienten entwickelt und ist sowohl in klinischen Studien als auch in der Routineversorgung weit verbreitet. Der Fragebogen umfasst insgesamt 30 Fragen zur Beurteilung der Lebensqualität von Krebspatienten multidimensional über 10 Subskalen. Der QLQ-C30 liefert somit differenzierte Daten über das Wohlbefinden des Patienten, die sowohl für die klinische Forschung als auch für die Versorgung relevant sind.

Seit seiner Einführung wurde der QLQ-C30 in mehr als 130 Sprachen übersetzt und in zahlreichen Kulturen und klinischen Kontexten validiert (13). Diese Internationalisierung ist von zentraler Bedeutung, da Krebs ein globales Gesundheitsproblem darstellt und vergleichbare Daten aus unterschiedlichen Ländern und Populationen benötigt werden, um universell anwendbare Erkenntnisse zu gewinnen. Der QLQ-C30, mittlerweile in der Version 3.0 wird heute weltweit in einer Vielzahl von Studien eingesetzt, von der Bewertung neuer Arzneimittel bis hin zu Forschungsarbeiten über die langfristigen Auswirkungen von Krebstherapien. Er ist der in Europa am häufigsten eingesetzte Fragebogen zur Lebensqualität für Krebspatienten (14).

1.3.2 MODULE

Obwohl der QLQ-C30 ein etabliertes Instrument zur Erfassung und Bewertung genereller gesundheitsbezogener Lebensqualitätsaspekte bei Krebspatienten ist, weist auch dieser Fragebogen gewisse Limitationen auf. So werden weder krankheitsspezifische Merkmale noch behandlungsbedingte Auswirkungen, wie z.B. das Auftreten bestimmter Symptome adäquat abgebildet. Um diesen Einschränkungen zu begegnen, wurde von der EORTC QLG ein modularer Ansatz entwickelt. Dieser sieht den QLQ-C30 als zentralen Kernfragebogen vor, der durch zusätzliche Module ergänzt werden kann, welche gezielt auf spezifische

Tumorentitäten, Therapieformen, Symptome oder besondere Dimensionen der Lebensqualität ausgerichtet sind (15).

Der Entwicklungsprozess der EORTC-Module gliedert sich dabei in vier Phasen (I – IV). Ab Abschluss von Phase III können Module bereits eingesetzt werden, obwohl sie noch nicht umfassend psychometrisch getestet wurden. Ihr Einsatz sollte daher mit entsprechender Vorsicht erfolgen. Module, die Phase IV abgeschlossen haben, gelten als vollständig validiert. Aktuell stellt die EORTC QLG 49 Module in unterschiedlichen Entwicklungsphasen bereit, davon 36 voll validierte Module zu einer Vielzahl onkologischer Entitäten. Zu den bekanntesten gehören etwa die Module für Brustkrebs (QLQ-BR42) oder Lungenkrebs (QLQ-LC29). Diese ermöglichen eine gezielte Erfassung spezifischer Problembereiche, die in der jeweiligen klinischen Situation relevant sind. Weitere Module befinden sich in Entwicklung, darunter z. B. das AYA-Modul für Jugendliche und junge Erwachsene oder das SURV-100-Modul zur Lebensqualität von Langzeitüberlebenden.

1.3.3 ITEM LIBRARY

Obwohl sich das Konzept eines modularen Fragebogensystems, bestehend aus dem QLQ-C30 als Kerninstrument und einem ergänzenden, auf die Tumorentität oder Studienpopulation zugeschnittenen Modul in der Praxis bewährt hat, treten insbesondere im Kontext der rasanten Entwicklung neuer onkologischer Therapien zunehmend Herausforderungen auf. Viele der bestehenden Module können die spezifischen Symptome und Probleme neuer Behandlungsformen nicht immer in ausreichendem Maße erfassen und abbilden. Da die Konzeption, Entwicklung und Validierung eines neuen Moduls mit erheblichem zeitlichem Aufwand von mehreren Jahren verbunden ist gelingt es nicht immer, die Vielfalt moderner Therapien zeitnah abzubilden. Auch für besondere Studienpopulationen, wie z.B. Patienten mit sehr seltenen Tumorerkrankungen stehen oftmals keine adäquaten Module zur Verfügung. In solchen Fällen hat sich gezeigt, dass ein flexiblerer und zugleich pragmatischer Ansatz erforderlich ist, um die gesundheitsbezogene Lebensqualität auch dieser Patientengruppen angemessen erfassen zu können.

Vor diesem Hintergrund hat die EORTC QLG die sogenannte Item Library als Ergänzung zum Kernfragebogen QLQ-C30 sowie zu bestehenden Modulen entwickelt und veröffentlicht. Die Item Library ermöglicht es Klinikern und Forschern, auf der

Basis einer Auswahl von über 1.000 entwickelten, getesteten und validierten Einzelfragen (sog. Items) gezielt die Symptome und Ereignisse zu erfassen, die für eine spezifische klinische Fragestellung von Relevanz sind. Damit bietet sie einen flexiblen Ansatz zur Erweiterung der bereits bestehenden Fragebögen (siehe Abbildung 1) (16). Zur Einhaltung methodischer Standards und einer standardisierten Anwendung wurde zudem ein eigenes User-Guideline-Dokument durch die EORTC QLG publiziert (17).

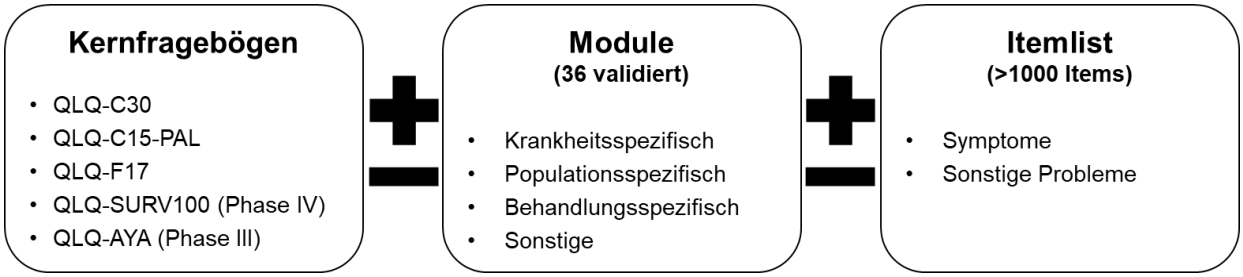


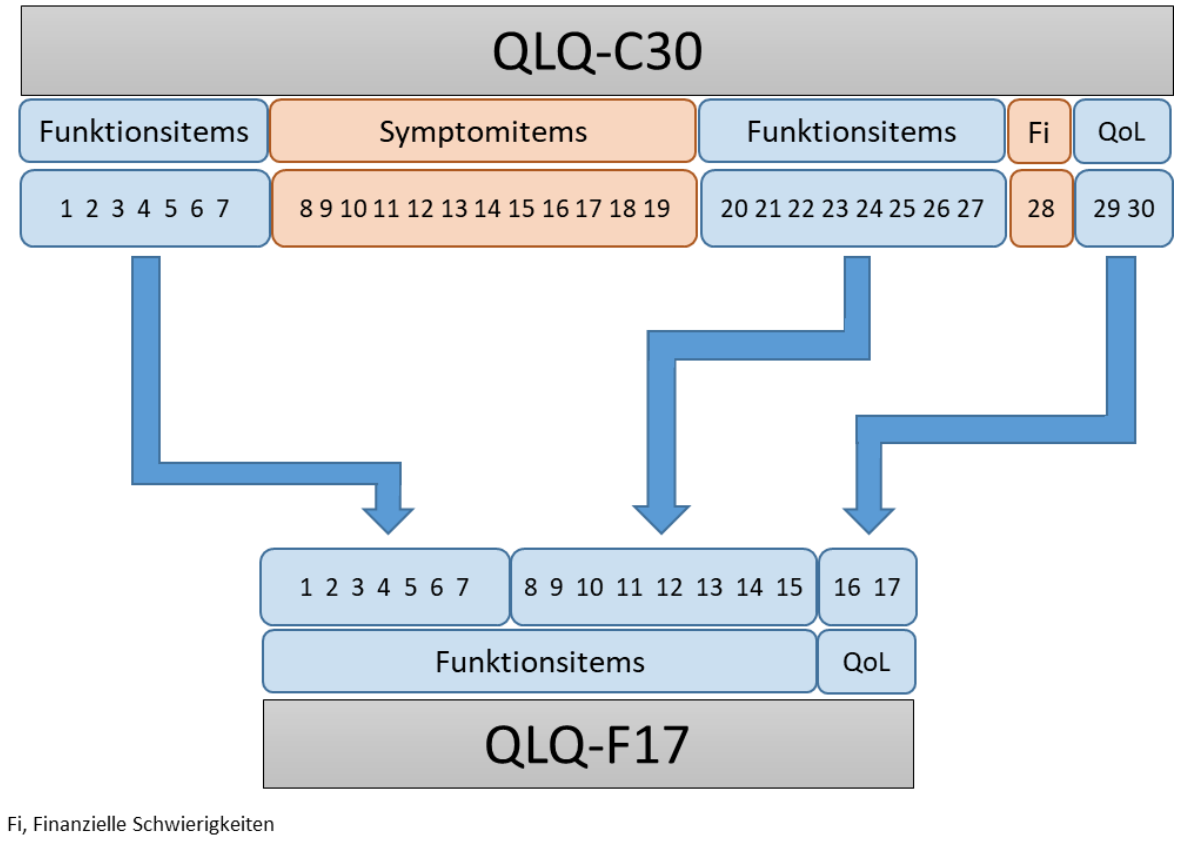
***Abbildung 1**: Modularer Ansatz der EORTC zur Erhebung von Lebensqualität*

1.3.4 EORTC QLQ-F17

In der klinischen Forschung kommen häufig umfangreiche Fragebögen zur Erhebung der Lebensqualität und des funktionalen Outcomes zum Einsatz. Regierungsbehörden wie die U.S. Food and Drug Administration (FDA) sowie Patienten selbst fordern jedoch zunehmend eine Reduktion und Straffung dieser Instrumente mit dem Ziel, die Belastung für Patienten, gerade in der belastenden Phase einer neu diagnostizierten Krebserkrankung und deren Therapie, möglichst gering zu halten (18). Ein Ziel muss demnach sein, die Anzahl und auch den Umfang der Fragebögen so zu gestalten, dass sie zum einen alle klinisch relevanten und notwendigen Informationen erfassen, dabei jedoch möglichst kurz und anwenderfreundlich bleiben.

Der EORTC QLQ-C30, als eines der etabliertesten Instrumente zur Erfassung der Lebensqualität in der Onkologie, enthält zahlreiche Fragen zu Symptomen, die im Kontext moderner, zielgerichteter oder immunonkologischer Therapien nicht immer angemessen erscheinen. Tatsächlich bezieht sich knapp die Hälfte der Fragen im QLQ-C30 auf Symptome, die in vielen aktuellen Therapieschemata eine untergeordnete Rolle spielen.

Als Reaktion auf diese Entwicklungen wurde von der EORTC QLG der QLQ-F17 entwickelt. Dieser Fragebogen ist eine gekürzte Version des QLQ-C30, welcher ausschließlich die Items zu den fünf Funktionsskalen sowie die beiden Items zur Skala zum globalen Gesundheitszustand/Lebensqualität umfasst (siehe Abbildung 2).



Fi, Finanzielle Schwierigkeiten

***Abbildung 2**: Aufbau des QLQ-F17*

Dabei verzichtet die 17-Item-Version bewusst auf die Symptomskalen des Originals mit dem Ziel einer gezielten und flexiblen Erfassung therapiespezifischer oder patientenrelevanter Symptome durch Ergänzung mit entsprechenden Modulen oder einzelnen Items aus der EORTC Item Library. Der QLQ-F17 stellt somit ein pragmatisches und adaptierbares Instrument dar, das den Anforderungen moderner, multimodaler Krebstherapien gerecht wird und gleichzeitig regulatorische sowie patientenzentrierte Erwartungen an die Reduktion der Erhebungsbelastung erfüllt.

Der QLQ-F17 kann somit als verkürzte Entsprechung zu den Funktionsskalen sowie zur Skala des globalen Gesundheitszustands/Lebensqualität des QLQ-C30 betrachtet werden. Durch die inhaltliche Übereinstimmung in diesen Bereichen ergibt sich die Möglichkeit, Studienergebnisse hinsichtlich der funktionalen Lebensqualität

unabhängig davon, ob der QLQ-C30 oder der QLQ-F17 als generisches Messinstrument verwendet wurde miteinander zu vergleichen. Der QLQ-F17 kann somit als ein neuer und weiterer Kernfragebogen betrachtet werden.

Allerdings wurde der QLQ-F17 bislang ausschließlich auf der Website der EORTC veröffentlicht, ohne dass bisher formale Publikationen zu seiner Validierung oder psychometrischen Prüfung vorliegen. Ebenso fehlt derzeit die empirische Evidenz, dass die Funktionsskalen in beiden Instrumenten tatsächlich vergleichbare Ergebnisse liefern. Für eine gleichwertige Nutzung beider Instrumente im wissenschaftlichen oder regulatorischen Kontext ist es jedoch essenziell, dass die erhobenen Werte unabhängig vom verwendeten Fragebogen äquivalent interpretierbar sind.

Auch wenn es naheliegend erscheint, anzunehmen, dass Skalenwerte zweier Fragebögen identisch ausfallen müssen, sofern dieselben Items verwendet werden, kann diese Äquivalenz im Fall des QLQ-C30 und des QLQ-F17 nicht ohne Weiteres vorausgesetzt werden. Der Grund dafür liegt in der unterschiedlichen inhaltlichen Struktur der Fragebögen. Im QLQ-C30 sind die Items zu den Funktionsskalen nicht in einem durchgängigen Block angeordnet, sondern werden durch symptombezogene Items unterbrochen. Dabei befinden sich die Symptomfragen an den Positionen 8 bis 19 und an der Position 28. Beim QLQ-F17 hingegen wurden diese symptom-bezogenen Items entfernt, wodurch sich die Struktur und somit der Kontext der verbleibenden Fragen verändert hat.

Forschungen zum Einfluss der Reihenfolge von Fragen auf Verzerrungen und Effekte innerhalb standardisierter Erhebungsinstrumente haben gezeigt, dass die Beantwortung einer Frage wesentlich durch die vorangegangenen Fragen beeinflusst sein kann (auch Reihenfolge-Effekte oder Positionseffekte genannt) (19),(20).

Insbesondere wurde nachgewiesen, dass vorangehende Fragen sowohl assimilative als auch kontrastive Effekte auf nachfolgende Antworten ausüben können. Assimilationseffekte treten dann auf, wenn sich Befragte bei der Beantwortung neuer Fragen an zuvor gegebenen Antworten orientieren und so eine konsistente Bewertung über zwei oder mehrere Fragen abgeben. Im Gegensatz dazu können Kontrasteffekte auftreten, bei denen abweichend geantwortet wird, um Redundanzen zu vermeiden oder Unterschiede zu betonen (21).

Beide Effekte können potenziell zu systematischen Verzerrungen führen, insbesondere dann, wenn Fragen innerhalb eines Instruments neu angeordnet oder entfernt werden. Daher ist es möglich, dass sich die veränderte Reihenfolge und reduzierte Itemstruktur des QLQ-F17 auf die Interpretation und Vergleichbarkeit der Funktionsskalen auswirkt. So könnte z.B. argumentiert werden, dass das Weglassen aller Symptom- und Finanzprobleme-Items des QLQ-C30 die Art und Weise verändert, wie die nachfolgenden Fragen beantwortet werden. Beispielsweise könnten Patienten, die ein hohes Maß an Atemnot und Schmerzen berichten, im Vergleich zu Patienten mit gleichem (objektivem) Gesundheitsstatus, denen diese Items nicht vorgelegt wurden, eine schlechtere emotionale Funktion angeben. Umgekehrt könnten Patienten, die keine Symptome in den Symptomskalen berichten, ihre Gesundheit als ausgezeichnet wahrnehmen und entsprechend eine bessere Funktion auf den Funktionsskalen angeben.

Eine empirische Überprüfung der Äquivalenz beider Versionen ist deshalb zwingend notwendig, bevor der QLQ-F17 als gleichwertige Alternative zum QLQ-C30 in wissenschaftlichen oder regulatorischen Kontexten verwendet werden kann.

1.4 ÄQUIVALENZSTUDIEN IN DER KLINISCHEN FORSCHUNG

Während das Konzept der Überlegenheitsstudie in der klinischen Forschung weit verbreitet und gut etabliert ist, werden Nicht-Unterlegenheits- und insbesondere Äquivalenzstudien vergleichsweise selten durchgeführt. Entsprechend ist die spezifische methodische Herangehensweise, die sich grundlegend von jener der Überlegenheitsstudie unterscheidet, oft wenig bekannt. Das Ziel einer Äquivalenzstudie besteht darin, nachzuweisen, dass eine neue Intervention im Hinblick auf den primären Endpunkt äquivalent zu einer etablierten Vergleichstherapie, in der Regel dem aktuellen Standard of Care ist. Äquivalenz lässt sich somit definieren, dass ein Patient keinen Unterschied in der Wirkung oder Sicherheit feststellen wird, wenn ein Medikament durch das andere ersetzt wird (22).

Ein grundlegender methodischer Unterschied zwischen einer Überlegenheitsstudie und einer Äquivalenzstudie liegt dabei in der Formulierung der statistischen Hypothesen. Während bei Überlegenheitsstudien die Nullhypothese von keinem Unterschied zwischen den Therapien ausgeht und die Studie darauf abzielt, diesen Gleichstand zugunsten eines Unterschieds (Alternativhypothese) zu widerlegen, ist es

bei Äquivalenzstudien genau umgekehrt. Hier unterstellt die Nullhypothese zunächst einen relevanten Unterschied zwischen den Behandlungen, und die Studie soll diesen Unterschied widerlegen, um daraus die Gleichwertigkeit im Sinne der Alternativhypothese ableiten zu können (siehe Abbildung 3).

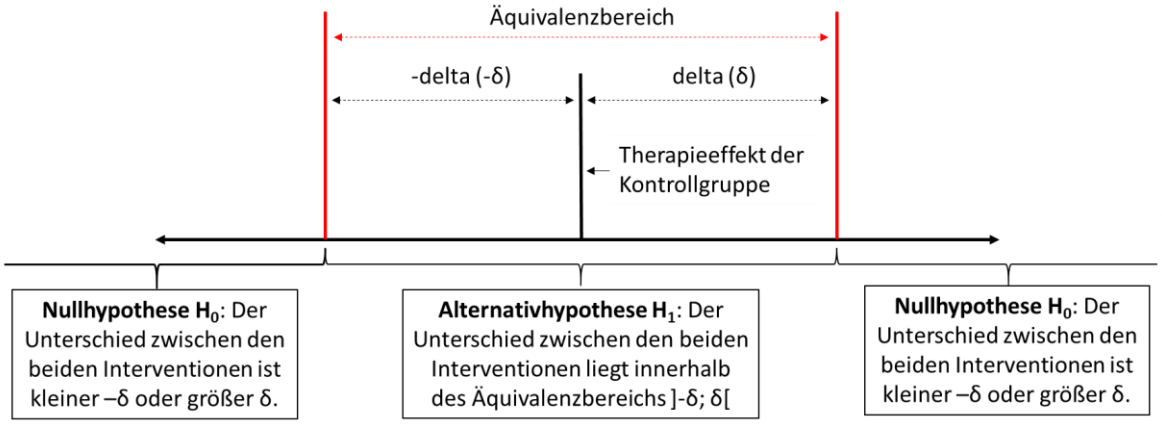


***Abbildung 3**: Hypothesenformulierung bei Äquivalenzstudien*

Eine der größten Herausforderungen bei der Planung von Äquivalenzstudien ist die Festlegung des sogenannten Äquivalenzbereichs (engl. equivalence margin). Für die Definition dieser Grenzen existieren methodische Empfehlungen und regulatorische Vorgaben, etwa von der US-amerikanischen Food and Drug Administration (FDA) sowie der Europäischen Arzneimittel-Agentur (EMA) (23),(24). Übereinstimmend wird empfohlen, dass die Festlegung des Äquivalenzbereichs sowohl auf statistischer Evidenz als auch auf klinischer Relevanz basieren sollte. Dieser Prozess ist von hoher Wichtigkeit, da die Definition des Äquivalenzbereichs die methodische Grundlage für die Interpretation der Studienergebnisse bildet. Eine weitere Herausforderung bei der Durchführung von Äquivalenzstudien stellt die benötigte Fallzahl dar. Da der statistische Nachweis der Äquivalenz voraussetzt, dass das Konfidenzintervall des geschätzten Effekts vollständig innerhalb der vorab definierten Äquivalenzgrenzen liegt, können gerade bei eng gewählten Äquivalenzbereichen schnell hohe Fallzahlen notwendig werden. Diese Anforderungen stellen sowohl in praktischer als auch in ethischer Hinsicht eine relevante Hürde bei der Planung und Durchführung entsprechender Studien dar.

1.5 ZIELE DER STUDIE

Primäres Ziel der vorliegenden Studie war es, zu untersuchen, ob der QLQ-F17 im Vergleich zum etablierten Kernfragebogen QLQ-C30 hinsichtlich der gemeinsamen Skalen (körperliche Funktion, Rollenfunktion, kognitive Funktion, emotionale Funktion, soziale Funktion und globaler Gesundheitszustand/Lebensqualität) äquivalente Ergebnisse liefert. Die Bewertung dieser Äquivalenz sollte dabei auf Grundlage zweier komplementärer methodischer Ansätze basieren. Zum einen war vorgesehen, die Übereinstimmung der Skalenwerte innerhalb vordefinierter Äquivalenzgrenzen zu überprüfen. Zum anderen sollte auf Item-Ebene analysiert werden, ob sich Antwortmuster bei vergleichbarer Lebensqualität nicht systematisch zwischen den beiden Fragebögen unterscheiden.

Sekundäre Zielsetzungen waren die Untersuchung potenzieller Reihenfolgeneffekte in weiteren Subgruppenanalysen sowie eine psychometrische Bewertung des QLQ-F17, insbesondere im Hinblick auf seine Reliabilität und Validität im Vergleich zum QLQ-C30.

Damit sollte die notwendige empirische Grundlage geschaffen werden, den QLQ-F17 künftig als alternativen Kernfragebogen zur Erfassung funktionaler Lebensqualität in onkologischen Studien einsetzen zu können.

2 MATERIAL UND METHODEN

2.1 STUDIENPLANUNG

2.1.1 SYSTEMATISCHE LITERATURRECHERCHE

Eine systematische Literaturrecherche dient dazu, den aktuellen Stand der Forschung zu einem spezifischen Thema umfassend zu erfassen und zu bewerten. Ziel der Literaturrecherche ist vorhandene Evidenz zu identifizieren, Wissenslücken aufzudecken und die Grundlage für weitere Forschungsarbeiten zu schaffen (25). Diese sollte stets bereits bei der Projektplanung durchgeführt werden.

Im Rahmen dieser Untersuchung wurde eine umfassende Literaturrecherche zur Validierung des QLQ-F17-Fragebogens in den Datenbanken PubMed, Embase, Web of Science Core Collection (SCI-EXPANDED, SSCI), Cochrane Library, ClinicalTrials.gov und Google Scholar durchgeführt. Dabei wurden sowohl wissenschaftliche Artikel als auch graue Literatur berücksichtigt, also Publikationen, die nicht über traditionelle Verlage veröffentlicht werden und oft schwerer zugänglich sind, wie z. B. Dissertationen oder Konferenzbeiträge. Lediglich ein Abstract konnte identifiziert werden, welcher den Fragebogen QLQ-F17 erstmalig vorstellt, jedoch keine weiteren Ergebnisse oder Evaluierungen beinhaltet. Auch eine erweiterte PubMed-Suche mit den Begriffen „functioning scales“ und „patient-reported outcomes“ führte zu lediglich 22 Publikationen, von denen nur eine Studie die Äquivalenz von Funktionsskalen mit einem anderen umfassenderen Langform-Fragebogen untersuchte. Die vorliegende Arbeit stellt somit eine Erstuntersuchung des QLQ-F17 dar, da bislang keine empirischen Untersuchungen zur Validierung oder Anwendung des Fragebogens verfügbar sind.

2.1.2 KONZEPT ZUR DATENERHEBUNG

Aufgrund des engen zeitlichen Rahmens des geförderten Projekts von zwei Jahren und der Zielsetzung, die Validierungsdaten zum QLQ-F17 zeitnah zu veröffentlichen, war eine Datenerhebung über die klassischen Rekrutierungswege der EORTC Quality of Life Group (QLG) nicht realisierbar. Feldstudien mit einer Zielgröße von 1.000 Patienten erfordern selbst bei Einbindung mehrerer nationaler EORTC Disease-Oriented Groups (DOGs) erfahrungsgemäß eine Laufzeit von über fünf Jahren. Für die

vorliegende Untersuchung mit einer angestrebten Fallzahl von rund 2.500 Patienten hätte sich die Rekrutierungsdauer entsprechend auf mindestens zehn Jahre verlängert.

Vor diesem Hintergrund und basierend auf früheren erfolgreichen Kooperationen mit kommerziellen Partnern in groß angelegten EORTC-Studien wurde die Datenerhebung an das internationale Marktforschungsunternehmen Kantar (www.kantar.com) ausgelagert (26). Kantar verfügt über mehr als 20 Jahre Erfahrung im Bereich der Gesundheitsforschung und ist weltweit in über 70 Ländern aktiv. Allein im Jahr 2019 führte das Unternehmen weltweit über 2.000 gesundheitsbezogene Studien durch. Kantar ist zudem nach der für die Markt-, Meinungs- und Sozialforschung international gültigen Norm ISO 20252:2019 zertifiziert. Die Erhebung der Daten erfolgt über validierte Patientenpanels mit Diagnosen aus verschiedenen Therapiegebieten, einschließlich onkologischer Erkrankungen in verschiedensten Entitäten. Kantar verfügt über eine umfangreiche Erfahrung in der Durchführung onkologischer Studien und war bereits an zahlreichen Projekten zur Lebensqualität, wie zum Beispiel der Europäischen Lebensqualitätsumfrage des Eurofound beteiligt (27).

Die Datenerhebung selbst erfolgt bei Kantar in der Regel internetbasiert mithilfe eines proprietären Erhebungstools. Teilnehmende werden aus bestehenden Patientenpanels rekrutiert und nehmen online an der Umfrage teil. Solche Befragungen können entweder in einer einzigen Sitzung abgeschlossen oder in mehreren Sitzungen über verschiedene Tage hinweg durchgeführt werden. Mehrteilige Erhebungen bieten zwar mehr Flexibilität, gehen jedoch häufig mit einem erhöhten Erhebungsaufwand und höheren Kosten einher. Zudem steigt das Risiko für vorzeitige Abbrüche (Drop-outs), was zu systematischen Verzerrungen der Ergebnisse führen kann. Aus diesen Gründen wurde bereits bei der initialen Antragstellung festgelegt, dass die Datenerhebung in einer einzigen Session erfolgen soll. Alle Teilnehmenden sollten den QLQ-C30 und den QLQ-F17 am selben Tag bearbeiten. Eine detailliertere Beschreibung zum Aufbau der Befragung findet sich in Kapitel *2.2.2 Studienaufbau*.

Zur Sicherstellung einer validen Durchführung der Datenerhebung wurden die EORTC-Leitlinien für elektronische Implementierungen von QOL-Instrumenten

berücksichtigt (28). Diese beinhalten Vorgaben zur originalgetreuen Übertragung papierbasierter Fragebögen in elektronische Formate. Dabei soll primär das Layout möglichst einfach und übersichtlich bleiben, und die inhaltliche Struktur unverändert übernommen werden. Weitere Aspekte umfassen zum Beispiel eine intuitive Navigation ohne erzwungene Antworten mit Möglichkeit zur Korrektur vor Abschluss, eine vollständige Darstellung von Antwortoptionen mit Text und Zahlen ohne Vorauswahl, konsistente und klare Instruktionen analog zur Papierfassung, keine Nutzung von Scroll-Funktionen, und viele weitere. Ziel der Leitlinien ist eine benutzerfreundliche, methodisch valide und regulatorisch konforme Umsetzung der EORTC-Fragebögen in elektronischen Anwendungen.

2.1.3 Entwicklung der Patientenbefragung

Für die Erstellung der Patientenbefragung wurde zunächst eine Draftversion des Gesamtfragebogens erstellt. Diese enthielt die beiden relevanten Fragebögen QLQ-C30 und QLQ-F17 in originaler Form, sowie weitere Fragen, welche zwischen den beiden Fragebögen eingefügt wurden. Die weiteren Fragen sollten zum einen patientenbezogene Informationen erheben (z. B. zum Gesundheitszustand, zur Behandlungssituation oder zum Alter und Geschlecht) als auch eine ablenkende Funktion haben, um die Erinnerungen an die Antworten zum zuerst bearbeiteten Fragebogen abzuschwächen. Die Ablenkungsfragen wurden dabei möglichst einfach gehalten und teils als Freitextfragen formuliert, um eine gedankliche Umstellung bei den Patienten anzustoßen. Die Bearbeitung der zusätzlichen Fragen wurde so konzipiert, dass sie eine Dauer von ca. 10 bis 15 Minuten beanspruchen sollten. Dies wurde als ausreichend erachtet, um einen zeitlichen Abstand zwischen den beiden Kernfragebögen herzustellen, ohne dabei die Akzeptanz oder Belastung der Patienten übermäßig zu erhöhen.

Nach der Entwicklung der ersten Draftversion des Gesamtfragebogens zur Erhebung der Studiendaten wurde dieser im Rahmen eines Reviews durch die Studiengruppe überprüft und unter Einbeziehung konzeptueller Empfehlungen seitens Kantars finalisiert. Besonderes Augenmerk galt dabei der inhaltlichen und formalen Gestaltung der zusätzlichen Fragen zwischen den beiden Fragebögen (QLQ-C30 und QLQ-F17), um potenzielle Beeinflussungseffekte systematisch zu minimieren. Die finale Fassung des Fragebogens wurde zunächst in deutscher Sprache erstellt und anschließend ins

Englische übersetzt, um eine konsistente Anwendung in beiden Sprachversionen sicherzustellen.

2.1.4 Vorstudie und Finalisierung der Patientenbefragung

Im nächsten Schritt der Studienvorbereitung wurden Pilotdaten am Universitätsklinikum Regensburg (UKR) in der onkologischen Tagesklinik erhoben, um die Verständlichkeit und Akzeptanz des entwickelten Fragebogens im klinischen Setting zu überprüfen. Insgesamt füllten acht Patienten den Fragebogen aus und gaben im Anschluss Rückmeldung zu Aufbau, Umfang und Verständlichkeit. Auf Basis der durchwegs positiven Rückmeldungen erfolgte die Finalisierung des Gesamtfragebogens in deutscher und englischer Sprache. Die anschließende Übersetzung in alle für die internationale Datenerhebung vorgesehenen Sprachversionen erfolgte durch Kantar. Für die Items des QLQ-C30 und des QLQ-F17 kamen dabei die bereits vorhandenen und validierten Übersetzungen gemäß den EORTC-Übersetzungsrichtlinien zum Einsatz (29). Nach Abschluss der Übersetzungsarbeiten wurden alle Sprachfassungen einer finalen inhaltlichen und formalen Überprüfung unterzogen.

2.1.5 Statistischer Analyseplan

Parallel zur Konzeption und Durchführung der Vorstudie wurde der statistische Analyseplan (SAP) erstellt und finalisiert. Der SAP bildet dabei die methodische Grundlage für die strukturierte und nachvollziehbare Auswertung der erhobenen Studiendaten (30). Ziel war es, bereits vor Beginn der Hauptanalyse klare Vorgaben für die statistischen Auswertungen zu formulieren und so die Transparenz, Reproduzierbarkeit und wissenschaftliche Qualität der Ergebnisse sicherzustellen. Der SAP umfasste unter anderem die Beschreibung der primären und sekundären Studienziele, die Definition der relevanten Studienpopulationen, den Umgang mit fehlenden oder unvollständigen Daten sowie die Auswahl und Begründung der eingesetzten statistischen Verfahren. Eine detaillierte Darstellung der im SAP definierten Methoden und Verfahren findet sich in Kapitel 2.6 *Statistische* Methoden.

2.2 STUDIENDURCHFÜHRUNG

Nach Abschluss der Studienplanung und der Finalisierung der Patientenbefragung wurde die Studie zur Erhebung der Studiendaten initiiert.

2.2.1 STUDIENDESIGN

Die Studie war als randomisierte, multinationale, fragebogenbasierte Cross-over-Studie konzipiert, mit dem Ziel, die Äquivalenz der funktionalen Skalen sowie der globalen Gesundheitszustands-/Lebensqualitätsskala zwischen dem QLQ-C30 und dem QLQ-F17 zu prüfen. Die Studie wurde bei der Ethikkommission der Universität Regensburg eingereicht und von dieser am 27. Juli 2022 (Referenznummer 22-3018-104) genehmigt. Im Sinne der Transparenz und Nachvollziehbarkeit wurde die Studie zudem im Primärregister ClinicalTrials.gov unter der Nummer NCT05479682 registriert.

Im Rahmen des Cross-over-Designs bearbeiteten alle Patienten beide Fragebögen zur Lebensqualität in einer einzigen Sitzung (siehe Kapitel *2.1.2 Konzept zur Datenerhebung*). Die Patienten wurden dabei zufällig einer von zwei Gruppen zugewiesen. In Gruppe 1 erfolgte die Bearbeitung der Fragebögen zur Lebensqualität in der Reihenfolge QLQ-C30 – Zwischenfragen - QLQ-F17, in Gruppe 2 in umgekehrter Reihenfolge (QLQ-F17 – Zwischenfragen - QLQ-C30). Die Zuteilung erfolgte zufällig und automatisiert nach dem Prinzip der geringsten Auslastung, um eine möglichst gleichmäßige Verteilung der Teilnehmenden auf beide Gruppen innerhalb jedes Landes sicherzustellen. Die Zuteilungsverschleierung (engl. allocation concealment) war gewährleistet, da die Zuordnung über die technische Infrastruktur der Umfrageplattform erfolgte und somit unabhängig von Eigenschaften der zuzuordnenden Patienten war.

2.2.2 STUDIENAUFBAU

Zu Beginn der Studie wurden die teilnehmenden Patienten darüber informiert, dass es sich um eine gesundheitsbezogene Befragung zum Thema Lebensqualität handelt. Die genaue Forschungsfragestellung wurde dabei nicht offengelegt, um eine möglichst unbeeinflusste Beantwortung der Fragebögen zu gewährleisten. In der ersten Frage mussten die Patienten zunächst ihr Einverständnis zur Teilnahme geben. Direkt im

Anschluss erfolgte eine Screeningfrage zur aktuellen oder zurückliegenden Krebserkrankung der Teilnehmenden:

Welche der folgenden Aussagen trifft am besten auf Sie zu:

1. *Bei mir wurde in den letzten 3 Monaten zum ersten Mal Krebs diagnostiziert.*
2. *Ich befinde mich derzeit in Krebstherapie.*
3. *Ich bin in Remission / Ich habe meinen Krebs besiegt.*
4. *Bei mir wurde noch nie Krebs diagnostiziert.*
5. *Anderes*

Patienten, die eine der Antwortoptionen 1, 2 oder 3 auswählten, konnten mit der Befragung fortfahren. Bei Auswahl von Option 4 oder 5 wurde die Teilnahme automatisch beendet, da diese Patienten nicht der Zielpopulation der Studie entsprachen.

Die Befragung selbst bestand aus drei Abschnitten. Als erste Erhebung wurde entweder der QLQ-C30 oder der QLQ-F17 je nach zufälliger Zuteilung im Rahmen des Cross-over-Designs ausgefüllt. Anschließend mussten die Patienten im Sinne einer Wash-out Phase eine Reihe von 34 unterschiedlichen Fragen beantworten. Diese Fragen sollten dazu beitragen, mögliche Erinnerungseffekte an den zuvor beantworteten Lebensqualitätsfragebogen zu minimieren. Eine detaillierte Beschreibung dieser Fragen findet sich in Kapitel *2.3.3 Zwischenfragen*. Nach Beantwortung der 34 Zwischenfragen folgte im dritten Abschnitt als zweite Erhebung die Bearbeitung des jeweils anderen Lebensqualitätsfragebogens (QLQ-C30 oder QLQ-F17), der im ersten Abschnitt noch nicht ausgefüllt worden war (siehe Abbildung 4).

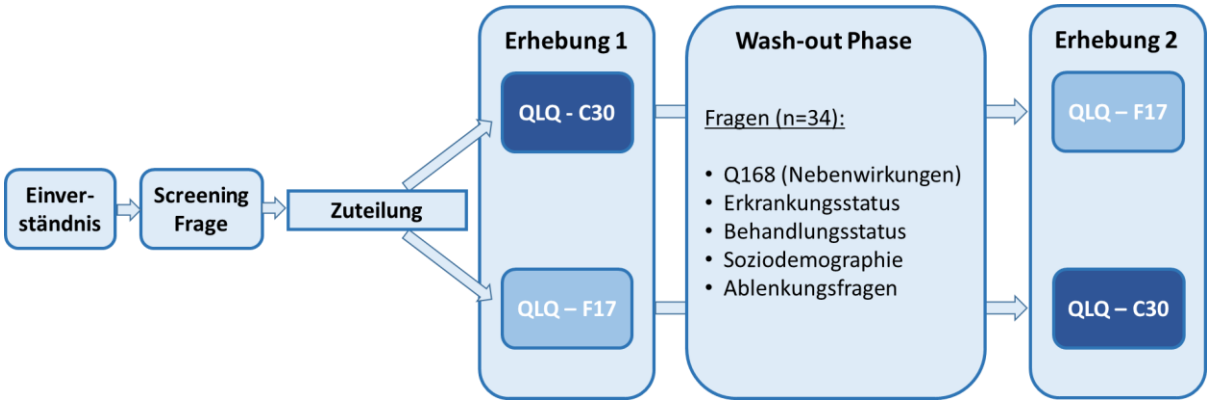


***Abbildung 4**: Übersicht Studienablauf (1)*

2.2.3 DATENERHEBUNG

Die Datenerhebung der Hauptstudie startete am 17.02.2023, wobei Patienten aus den 11 Ländern Australien, Finnland, Frankreich, Deutschland, Italien, Polen, Rumänien, Spanien, Schweden, Vereinigtes Königreich und USA rekrutiert wurden. Basierend auf der Einwohnerzahl und der damit verbundenen Anzahl von verfügbaren Patienten, wurde für jedes Land ein eigenes Rekrutierungsziel definiert, um die angestrebte Patientenzahl von mindestens n=2.500 Patienten (siehe Kapitel *2.5 Fallzahlberechnung*) zu erreichen. Als weiteres Stratum wurde zusätzlich noch das Geschlecht eingeführt um eine annähernde Gleichverteilung des Geschlechts innerhalb jedes Landes zu erreichen. Um Abbrecher und Patienten mit schlechter Datenqualität zu kompensieren, wurde die Zielpopulation von Kantar etwas erhöht. Das finale Rekrutierungsziel je Land ist in Tabelle 1 dargestellt.

Um frühzeitig auf systematische Fehler reagieren zu können, und um die Qualität der bis dahin erhobenen Daten zu überprüfen, wurde die Rekrutierung langsam gestartet und nach 10% der rekrutierten Patienten eine Rekrutierungspause eingelegt. Dabei wurde überprüft, ob die Dauer der Fragebögen in einem zu erwartenden Maß war, wie hoch die Quote der frühzeitigen Abbrecher war und wie viele Screeningfehler es in den einzelnen Ländern gab. Am 28.03.2023 wurden alle Rekrutierungsziele der einzelnen Länder erreicht und die Datenerhebung abgeschlossen.

***Tabelle 1**: Rekrutierungsziele*

Land	Ziel-population
Australien	150
Deutschland	300
Finnland	150
Frankreich	300
Italien	300
Polen	250
Rumänien	150
Schweden	150
Spanien	300
UK	200
USA	420
Gesamt	**2670**

2.3 BESCHREIBUNG DER MESSINSTRUMENTE

2.3.1 EORTC QLQ-C30 (VERSION 3.0)

Mit dem EORTC QLQ-C30 als Kernfragebogen werden allgemeine Aspekte der Lebensqualität von Krebspatienten während der letzten Woche multidimensional über unterschiedliche Skalen erfasst. Er umfasst 30 Fragen (Items) welche sich in fünf funktionale Skalen, einer Skala zum Globalen Gesundheitszustand/Lebensqualität (LQ), drei Symptomskalen sowie sechs Einzelitems aufteilen. Alle Items außer die beiden Items der LQ-Skala werden auf einer 4-Punkte-Likert-Skala mit den Antwortmöglichkeiten „überhaupt nicht“, „wenig“, „mäßig“ und „sehr“ bewertet. Die beiden Items der LQ-Skala werden auf einer 7-Punkte-Skala von 1 (sehr schlecht) bis 7 (ausgezeichnet) bewertet (siehe Tabelle 2). Der Fragebogen findet sich in seiner Originalform in deutscher Übersetzung im Anhang A.

***Tabelle 2**: Aufbau und Skalen des EORTC QLQ-C30*

Übergeordnete Skalengruppe	Skala (Abkürzung*)	Item #	Anzahl Items	Antwort-skala
Funktionale Skalen	Körperliche Funktion (PF)	1–5	5	4-Punkte
	Rollenfunktion (RF)	6, 7	2	4-Punkte
	Kognitive Funktion (CF)	21–24	4	4-Punkte
	Emotionale Funktion (EF)	20, 25	2	4-Punkte
	Soziale Funktion (SF)	26, 27	2	4-Punkte
Symptom-skalen	Müdigkeit (FA)	10, 12, 18	3	4-Punkte
	Schmerzen (PA)	9, 19	2	4-Punkte
	Übelkeit/Erbrechen (NV)	14, 15	2	4-Punkte
Symptom-Einzelitems	Dyspnoe (DY)	8	1	4-Punkte
	Appetitlosigkeit (AP)	13	1	4-Punkte
	Schlaflosigkeit (SL)	11	1	4-Punkte
	Verstopfung (CO)	16	1	4-Punkte
	Durchfall (DI)	17	1	4-Punkte
	Finanzielle Schwierigkeiten (FI)	28	1	4-Punkte
Allgemeine Lebensqualität	Globaler Gesundheitszustand / Lebensqualität (QL)	29, 30	2	7-Punkte

**Die Abkürzungen basieren auf den englischen Bezeichnungen der Skalen; 4-Punkte: Likertskala „überhaupt nicht“, „wenig“, „mäßig“ und „sehr“; 7-Punkte: Skala von 1 (sehr schlecht) bis 7 (ausgezeichnet)*

2.3.2 EORTC QLQ-F17

Der QLQ-F17 ist eine verkürzte Version des QLQ-C30, bei der alle Skalen und Einzelitems zu Symptomen und finanziellen Schwierigkeiten entfernt wurden. Der 17-teilige Fragebogen umfasst daher nur die fünf Funktionsskalen (PF, RF, CF, EF und SF) und die Skala zum globalen Gesundheitszustand/Lebensqualität (siehe *1.3.4 EORTC QLQ-F17)*.

2.3.3 Zwischenfragen

Zwischen den beiden Fragebogenversionen wurde ein zusätzlicher Fragenblock eingeschoben, der sowohl medizinisch relevante Informationen als auch soziodemografische und alltagsbezogene Aspekte erfasst. Dazu zählen Fragen zur Tumorart, zum Zeitpunkt der Diagnose, aktuellen Behandlungsformen und etwaigen Begleiterkrankungen sowie zum subjektiven Allgemeinbefinden und Aktivitätsgrad. Ergänzend wurden Angaben zu Alter, Geschlecht, Bildung, Berufstätigkeit, Lebenssituation und sportlichen Aktivitäten erhoben. Darüber hinaus umfasste der Fragenblock auch allgemeinere Themen des persönlichen Lebensstils und der Freizeitgestaltung, darunter bevorzugte Hobbys, Urlaubsarten, Medien- und Musikkonsum, Essgewohnheiten, Lieblingstiere, Automarken, Sprachen und IT-Nutzung. Ziel dieser Zwischenfragen war es, den Übergang zwischen den beiden funktionalen Fragebogenteilen methodisch aufzulockern und potenzielle Reihenfolgeeffekte in der Beantwortung zu entschärfen. Eine vollständige Liste aller in der Befragung enthaltenen Fragen befindet sich im Anhang B.

2.4 Methodische Definitionen

2.4.1 Äquivalenzgrenzen

Für die statistische Bewertung der Äquivalenz zwischen zwei Verfahren ist die Festlegung eines inhaltlich begründeten Äquivalenzbereichs erforderlich. Nur innerhalb dieses Bereichs können Unterschiede als klinisch nicht relevant und die Verfahren somit als gleichwertig betrachtet werden. Wie in Kapitel *1.4 Äquivalenzstudien in der klinischen Forschung* dargelegt, ist diese Definition eine zentrale Voraussetzung für die Durchführung und Interpretation von Äquivalenzanalysen in der klinischen Forschung.

Die Definition geeigneter Äquivalenzgrenzen basiert dabei sowohl auf der klinischen Relevanz aus Patientensicht als auch auf der Einschätzung fachlicher Experten. Ein zentrales Konzept in diesem Zusammenhang ist die sogenannte minimale klinisch relevante Differenz (engl. minimal clinically important difference, MCID). Die MCID beschreibt die kleinste Veränderung eines Messergebnisses, die von Patienten und/oder Behandlern als relevant wahrgenommen wird (31).

Für die Skalen des QLQ-C30 wurde die MCID in den vergangenen 15 Jahren in zahlreichen Studien untersucht und diskutiert. In einer ersten Veröffentlichung zu diesem Thema schlugen Osoba et al. Grenzwerte von 5 bis 10 Punkten für ausgewählte Funktionsskalen (körperlich, emotional, sozial) sowie die Skala zur globalen Lebensqualität vor (32). Eine vertiefende Literaturrecherche zusammen mit Expertenmeinungen wurde später von Cocks et al. publiziert (33). Weitere MCID-Schätzungen basieren auf Anker-basierten Methoden und wurden von der EORTC Quality of Life Group in verschiedenen Tumorentitäten durchgeführt (34)–(43).

Eine kürzlich von Musoro et al. publizierte Synthese, basierend auf 21 klinischen Studien, zeigt, dass die MCID-Werte je nach Skala, Krebsart und Richtung der Veränderung (Verbesserung vs. Verschlechterung) zwischen 4 und 15 Punkten variieren können. Dennoch liegen die meisten MCID-Schätzungen für die Funktionsskalen im Bereich zwischen 5 und 10 Punkten (44).

Vor dem Hintergrund dieser Evidenz sowie in enger Abstimmung mit dem statistischen Advisory Board, bestehend aus Experten aus dem Bereich der Lebensqualitätsforschung der EORTC wurde für die vorliegende Studie ein konservativer, über alle Skalen einheitlicher Äquivalenzbereich von –5 bis +5 Punkten definiert.

2.4.2 Analysepopulation und Datenqualität

Bereits vor der Auswertung einer klinischen Studie ist eine klare Definition der Analysepopulation von zentraler Bedeutung. Diese legt fest, welche Patienten in die finale Analyse einbezogen werden und welche aus dem Datensatz ausgeschlossen werden müssen. Die Gründe für einen Ausschluss können vielfältig sein und sind abhängig vom Studiendesign unterschiedlich zu bewerten. Dabei besteht insbesondere bei onlinebasierten Umfragen grundsätzlich ein erhöhtes Risiko für eine eingeschränkte Datenqualität einzelner Teilnehmer, die potenziell zu Verzerrungen der Ergebnisse führen können. Anders als in kontrollierten klinischen Settings lässt sich der Befragungsprozess in webbasierten Studien nur bedingt überwachen. Dies betrifft sowohl den Kontext der Teilnahme (z. B. Ablenkung durch Umgebung oder parallele Tätigkeiten) als auch das Antwortverhalten der Teilnehmer. In der wissenschaftlichen Literatur wurden bereits zahlreiche Arbeiten publiziert, die sich mit

den Herausforderungen und Ursachen eingeschränkter Datenqualität in Online-Umfragen befassen und konkrete Kriterien zur Identifikation minderwertiger Datensätze vorschlagen (45). Faktoren wie die Art des verwendeten Endgeräts, die Länge und Komplexität des Fragebogens, das Interesse am Befragungsthema oder monetäre Anreize zur Teilnahme können das Antwortverhalten maßgeblich beeinflussen. Zur objektiven Identifikation unzuverlässiger Datensätze und zur Beurteilung minderwertige Datensätze sind mittlerweile viele Umfragemetriken definiert, wie zum Beispiel Fragebogen-Abbruchraten (*questionnaire completion rates*), Bearbeitungszeiten auf Fragebogen- und Item-Ebene (*completion times, item response times*), Antwortqualität und -genauigkeit (*response quality/accuracy*) und Konsistenzindikatoren und Plausibilitätsprüfungen (46).

Vor diesem Hintergrund wurde im Rahmen der vorliegenden Studie ein systematischer Ansatz zur Sicherstellung der Datenqualität entwickelt, um die Analysepopulation methodisch fundiert zu definieren. Zunächst wurde die Qualität jeder einzelnen Befragung überprüft, um unplausible oder unzureichende Datensätze vorab auszuschließen. Basierend auf der vorhandenen Literatur und den im Rahmen der Studie verfügbaren Kennwerten wurde ein objektivierbares Punktesystem implementiert, das auf vordefinierten Qualitätsindikatoren beruhte. Für jeden nicht erfüllten Qualitätsindikator wurde ein Punkt („Flag“) vergeben, wobei maximal 10 Punkte erreicht werden konnten. Patienten mit 0 bis maximal 2 Punkte galten als qualitativ ausreichend und wurden in die Analysepopulation aufgenommen. Datensätze mit 3 oder mehr Punkten wurden als unzureichend eingestuft und ausgeschlossen. Die genauen Definitionen der verwendeten Qualitätsindikatoren sind im Folgenden aufgeführt.

2.4.2.1 QUALITÄTSINDIKATOR: KONSISTENZ DER ERSTEN 7 FRAGEN

Ein zentrales Qualitätskriterium in dieser Studie betraf die Konsistenz der Antworten auf die ersten sieben Fragen der beiden Fragebögen QLQ-C30 und QLQ-F17. Diese sieben Items sind inhaltlich und formal identisch, erscheinen jeweils am Anfang des jeweiligen Fragebogens und werden auf einer vierstufigen Likert-Skala („überhaupt nicht“, „wenig“, „ziemlich“, „sehr“) beantwortet. Da die Fragen innerhalb einer einzigen Sitzung gestellt wurden, ist unter Annahme eines konsistenten Antwortverhaltens eine weitgehende Übereinstimmung der Angaben innerhalb eines Patienten zu erwarten.

Zur Bewertung der Konsistenz wurde überprüft, ob zwischen den Antworten auf ein identisches Item in beiden Fragebögen eine maximale Abweichung bestand, d. h. eine Antwortkombination von „überhaupt nicht“ und „sehr“ oder umgekehrt. Solche diametralen Diskrepanzen wurden als Hinweise auf unzuverlässiges Antwortverhalten interpretiert. Für jede der sieben Fragen wurde für eine maximale Abweichung ein Punkt vergeben, sodass innerhalb dieses Qualitätsindikators bis zu sieben Punkte erreicht werden konnten.

Diese Metrik stellt ein sensibles und zugleich differenziertes Maß zur Identifikation von careless responding im Rahmen eines Cross-over-Designs dar. Sie erlaubt es, stark inkonsistente Antwortmuster zu erkennen, ohne vereinzelte Abweichungen überzubewerten. Nur wenn mehrere stark abweichende Angaben auftraten, wurde dies als Indiz für eingeschränkte Datenqualität gewertet und entsprechend berücksichtigt.

2.4.2.2 Qualitätsindikator: Plausibilität der Antwortzeiten

Der zweite zentrale Aspekt der Qualitätsprüfung betraf die Analyse der Antwortzeiten innerhalb der Befragung. Im Rahmen der Online-Erhebung wurden sowohl die Gesamtbearbeitungszeit der Befragung als auch die Bearbeitungsdauer einzelner Abschnitte, insbesondere der beiden Lebensqualitätsfragebögen QLQ-C30 und QLQ-F17, systematisch erfasst. Unplausibel kurze Antwortzeiten können auf ein unaufmerksames oder mechanisches Antwortverhalten hinweisen und wurden daher gezielt als potenzielles Qualitätsproblem bewertet. Ziel dieses Qualitätsindikators war insbesondere die Identifikation sogenannter *Speeder*, also Teilnehmender, die den Fragebogen ungewöhnlich schnell und potenziell ohne inhaltliche Auseinandersetzung mit den Fragen ausfüllen.

Zur Identifikation solcher Muster wurden drei Schwellenwerte definiert, deren Unterschreitung jeweils als Hinweis auf mangelnde Sorgfalt gewertet wurde.

- Die Gesamtdauer der Befragung lag unterhalb der Hälfte des Medians der gesamten Stichprobe.
- Die durchschnittliche Bearbeitungszeit pro Item des QLQ-C30 lag unter 2 Sekunden.

- Die durchschnittliche Bearbeitungszeit pro Item des QLQ-F17 lag unter 2 Sekunden.

Für jede erfüllte Bedingung wurde ein weiterer Punkt vergeben, sodass innerhalb dieses Qualitätsindikators bis zu drei zusätzliche Punkte möglich waren. Die Kombination dieser drei Kennzahlen ermöglicht eine differenzierte Einschätzung der Antwortplausibilität, sowohl auf Ebene der Gesamtumfrage als auch bezogen auf die zentralen Studieninstrumente. Eine auffällige zu kurze Beantwortungszeit in einem oder mehreren dieser Bereiche wurde als Indiz für eingeschränkte Datenqualität berücksichtigt.

2.5 FALLZAHLBERECHNUNG

Eine sorgfältige Fallzahlplanung ist ein zentraler Bestandteil jeder klinischen Studie. Sie stellt sicher, dass die Anzahl der eingeschlossenen Patienten ausreichend ist, um das primäre Studienziel mit der gewünschten statistischen Aussagekraft zu überprüfen. Eine zu geringe Fallzahl birgt das Risiko, wahre Effekte nicht erkennen zu können (unzureichende Power), während eine zu große Fallzahl unnötigen Aufwand und Belastung für Patienten und Ressourcen bedeutet. Die Fallzahlberechnung erfolgt daher in Abhängigkeit von Ziel, Design und statistischer Methodik der Studie.

Die Berechnung der notwendigen Fallzahl der vorliegenden Studie basierte auf dem primären Studienziel, dem Nachweis der Äquivalenz der funktionalen Skalen sowie der Skala zum globalen Gesundheitszustand/ Lebensqualität zwischen dem QLQ-F17 und dem QLQ-C30. Grundlage der Berechnung war der Vergleich der Skalenwerte aus der ersten Erhebung zwischen zwei unabhängigen Patientengruppen, also der Vergleich der beiden Fragebogenversionen in einer Parallelgruppenstruktur.

Diese Herangehensweise beruhte auf zwei zentralen Überlegungen. Erstens erfordert der Vergleich zwischen unabhängigen Gruppen immer eine deutlich größere Fallzahl als der gepaarte Vergleich innerhalb derselben Gruppe, um dieselbe Teststärke (Power) zu erreichen. Wäre die Fallzahlplanung auf einem gepaarten Design durchgeführt worden, hätte dies bei einem späteren Gruppenvergleich eine zu geringe Power bedeutet. Zweitens war vor Studienbeginn nicht sicherzustellen, dass der gewählte Abstand zwischen den beiden Erhebung des QLQ-C30 und des QLQ-F17, die sogenannte Wash-out Phase, ausreichend lange gewählt war, um mögliche Carry-

over-Effekte auszuschließen. Um daher eine valide und sichere Aussagekraft mit der geplanten Power von mindestens 80 % zu gewährleisten, wurde die Fallzahl konservativ auf Basis des Vergleichs der ersten Erhebung zwischen zwei unabhängigen Gruppen berechnet.

Grundlage für die Berechnung der erforderlichen Fallzahl war eine Annahme zu den zu erwartenden Unterschieden zwischen dem QLQ-C30 und dem QLQ-F17 für alle funktionalen Skalen sowie der Skala zum globalen Gesundheitszustand/ Lebensqualität. Da im Rahmen einer Äquivalenzprüfung grundsätzlich davon ausgegangen wird, dass keine klinisch relevanten Unterschiede zwischen den zu vergleichenden Methoden bestehen, ist für die Fallzahlplanung dennoch eine quantitative Annahme zur erwarteten Differenz erforderlich. Um das Risiko einer unzureichenden Power für jede der Skalen zu minimieren, wurden die Annahmen zu den erwartenden Effekten bewusst konservativ gewählt. Für die mittlere Differenz zwischen den Skalenwerten des QLQ-C30 und des QLQ-F17 wurde eine maximale Abweichung von ±2 Punkten angenommen. Die zu erwartende Standardabweichung (SD) wurde basierend auf der eigenen Erfahrung sowie auf dem Referenzmanual der EORTC ebenso konservativ mit SD=30 für jede Skala geschätzt (47). Der Äquivalenzbereich wurde wie unter Kapitel *2.4.1 Äquivalenzgrenzen* beschrieben für alle Skalen von]-5, 5[festgelegt. Weiterhin wurde das zweiseitige Signifikanzniveau für alle Tests auf α=0,05 festgesetzt und die angestrebte Power auf mindestens 80% (Fehler 2. Art maximal β=0,2). Basierend auf diesen Annahmen ergab sich ein erforderlicher Stichprobenumfang von ca. n = 2.500 Patienten, entsprechend einem n = 1.250 pro Fragebogen innerhalb der ersten Erhebung. Diese Fallzahl erlaubte es, die Nullhypothese der Nicht-Äquivalenz unter Annahme eines Unterschieds von ±2 Punkten, einer Standardabweichung von 30 und eines Äquivalenzbereichs von]-5, 5[zu einem zweiseitigen Signifikanzniveau von 0,05 mit mindestens 80% Power zu verwerfen. Die Berechnung des Stichprobenumfangs wurde mit SAS v9.4 (SAS Institute, Cary, NC) durchgeführt (siehe Abbildung 5).

Das SAS System

The POWER Procedure
Equivalence Test for Mean Difference

Fixed Scenario Elements	
Distribution	Normal
Method	Exact
Lower Equivalence Bound	-5
Upper Equivalence Bound	5
Mean Difference	2
Standard Deviation	30
Nominal Power	0.8
Alpha	0.05
Group 1 Weight	1
Group 2 Weight	1

Computed N Total	
Actual Power	N Total
0.800	2476

***Abbildung 5**: Fallzahlberechnung mit SAS 9.4*

2.6 Statistische Methoden

Alle statistischen Analysen wurden mit der Statistiksoftware R (Version 4.3.2; R Core Team 2024) durchgeführt. In den finalen Datensatz gingen ausschließlich Daten von Patienten ein, die die Studie nicht vorzeitig abgebrochen haben. Diese Patienten haben beide Lebensqualitätsfragebögen (QLQ-C30 bzw. QLQ-F17) sowie die relevanten Zwischenfragen vollständig ausgefüllt, sodass für die analysierten Variablen keine fehlenden Werte vorlagen. Entsprechend war der Einsatz von Methoden zum Umgang mit fehlenden Werten nicht erforderlich.

2.6.1 Deskriptive Statistiken

Die Patientenmerkmale wurden für alle Patienten zusammengefasst und nach der Reihenfolge der präsentierten Fragebögen gruppiert (QLQ-C30 – QLQ-F17 vs. QLQ-F17 – QLQ-C30). Für die Ergebnisdarstellung wurden kontinuierliche Variablen als Mittelwert, Standardabweichung (SD), Minimum und Maximum berichtet. Kategoriale Variablen wurden durch absolute Häufigkeiten und Prozentsätze dargestellt.

2.6.2 Berechnung der Skalenscores des QLQ-C30 und des QLQ-F17

Die Berechnung der Skalenwerte für den QLQ-C30 und den QLQ-F17 erfolgte gemäß den Vorgaben des EORTC QLQ-C30 Scoring Manuals (48). Zunächst wurde für jede Skala der Rohwert (Raw Score, RS) als arithmetisches Mittel der zugehörigen Items ermittelt:

$$Rohscore\ (RS) = \frac{I_1 + I_2 + \cdots + I_n}{n}$$

Anschließend wurden die Rohwerte durch eine lineare Transformation in standardisierte Skalenwerte überführt, die auf einer einheitlichen Skala von 0 bis 100 liegen. Dabei wurde die unterschiedliche Codierungsrichtung der Skalen berücksichtigt. Während die Items der Funktionsskalen bei niedrigen Werten eine bessere Funktion zeigen, deuten höhere Werte bei den Items der Symptomskalen auf eine stärkere Beeinträchtigung hin. Ebenso bedeuten höhere Werte bei der Beurteilung des globalen Gesundheitszustands/Lebensqualität auch eine bessere Lebensqualität.

Im Zuge der Transformation wurden die Funktionsskalen invertiert, sodass alle resultierenden Skalenwerte einheitlich interpretierbar waren:

- Höhere Werte in den Funktionsskalen sowie der globalen Lebensqualität stehen für ein besseres Funktionsniveau bzw. höhere Lebensqualität.
- Höhere Werte in den Symptomskalen sowie bei einzelnen Symptombezogenen Items spiegeln eine stärkere Beeinträchtigung oder Symptomlast wider.

Die Transformation erfolgte nach den folgenden Formeln:

Für die Funktionsskalen:

$$Score = \left(1 - \frac{RS - 1}{range}\right) * 100$$

Für die Symptomskalen und des globalen Gesundheitszustands/Lebensqualität:

$$Score = \left(\frac{RS - 1}{range}\right) * 100$$

Dabei entspricht die *range* der Differenz zwischen dem maximalen und minimalen möglichen Rohwert der jeweiligen Skala.

2.6.3 Überprüfung der Äquivalenz

Zur Beantwortung der primären Fragestellung dieser Arbeit, dem Nachweis der Äquivalenz der gemeinsamen Skalen des QLQ-F17 und des QLQ-C30 wurden zwei methodische Ansätze verfolgt. Im ersten Ansatz erfolgte ein Vergleich zwischen zwei unabhängigen Gruppen (Zwischen-Gruppenvergleich) entsprechend der Rationalen der Fallzahlplanung. Dabei wurden ausschließlich der Ergebnisse aus der ersten Erhebung der jeweiligen Fragebögen herangezogen (QLQ-C30 aus Erhebung 1 vs. QLQ-F17 aus Erhebung 1). Im zweiten Ansatz wurden die wiederholten (gepaarten) Messungen beider Fragebögen innerhalb des Gesamtkollektivs analysiert.

Da bei den ersten sieben Items, die in beiden Fragebögen an gleicher Position abgefragt wurden und die körperliche Funktion und die Rollenfunktion erfassen, keine Reihenfolgeeffekte zu erwarten waren, lag der primäre Fokus der Analysen auf den verbleibenden zehn Items. Diese erfassen die emotionale Funktion, kognitive Funktion, soziale Funktion sowie den globalen Gesundheitszustand bzw. die Lebensqualität. Die Unterschiede in den ersten sieben Items bzw. den zugehörigen zwei Skalen wurden als zufällige Fehlerquellen interpretiert und dienten sowohl in den Analysen zwischen den Gruppen als auch innerhalb der Gruppen als Referenzwerte zur Einordnung der Effekte der verbleibenden zehn Items.

Ein erfolgreicher Nachweis der Äquivalenz wurde dann als erreicht gewertet, wenn sowohl die Zwischen-Gruppenvergleiche als auch die Innerhalb-Gruppenvergleiche konsistente und signifikante Ergebnisse lieferten. Nur wenn beide methodischen Ansätze zu übereinstimmenden Ergebnissen führten, wurde von einer robusten Evidenz für die Äquivalenz der jeweiligen Skalen ausgegangen.

2.6.3.1 Zwischen-Gruppenvergleiche

Auf Basis der ersten Erhebung (Zwischen-Gruppenvergleiche) wurden zur Überprüfung der Äquivalenz zwischen dem QLQ-C30 und dem QLQ-F17 zwei methodische Ansätze verwendet. Zum einen wurde das Differential Item Functioning (DIF) herangezogen, um potenzielle Unterschiede zwischen den Fragebögen auf

Ebene einzelner Items zu identifizieren. Zum anderen wurden multiple lineare Regressionsmodelle eingesetzt, um die Äquivalenz auf Skalenebene zu analysieren. Im Folgenden werden beide Ansätze sowohl inhaltlich als auch mathematisch näher erläutert.

Differential Item Functioning

Differential Item Functioning (DIF) ist ein Konzept aus der psychologischen Testtheorie und beschreibt den Umstand, dass Personen aus verschiedenen Gruppen, wie z.B. hinsichtlich Geschlecht, Alter, Sprache oder kultureller Zugehörigkeit bei einer bestimmten Frage unterschiedliche Antwortwahrscheinlichkeiten aufweisen, obwohl sie in der zugrunde liegenden Persönlichkeitseigenschaft bzw. dem sogenannten latenten Kontstrukt (z. B. der Lebensqualität) denselben Ausprägungsgrad aufweisen (49).

DIF liegt beispielsweise dann vor, wenn Befragte aus unterschiedlichen Ländern ein Item unterschiedlich beantworten, obwohl sie in Bezug auf das zu messende Merkmal vergleichbar sind. Dies kann zu Verzerrungen in den Testergebnissen führen, da der Vergleich zwischen Gruppen nicht aufgrund tatsächlicher Unterschiede im latenten Konstrukt, sondern aufgrund unterschiedlicher Reaktionen auf bestimmte Formulierungen oder kulturelle Interpretationen eines Items unterschiedlich ausfällt. Man spricht in diesem Fall von einer Gruppenabhängigkeit des Antwortverhaltens.

Zwei anschauliche Beispiele für die Anwendung von DIF-Analysen im Rahmen des EORTC QLQ-C30 finden sich in zwei publizierten Studien von Scott (50),(51). In seiner ersten Publikation aus dem Jahr 2006 analysierten Scott und Kollegen die Antworten von 27.891 Patienten aus 103 Studien, um sprachbedingte Unterschiede in der Beantwortung des QLQ-C30 zu identifizieren (50). Dabei wurde mit Hilfe von DIF geprüft, ob einzelne Items, nach Kontrolle für die zugrundeliegende Subskala, je nach Sprachversion unterschiedlich beantwortet wurden. Für jede der 13 untersuchten Übersetzungen wurde mindestens ein Item mit signifikantem DIF gefunden. In einigen Fällen waren die Unterschiede so stark, dass sie potenziell die Ergebnisse multinationaler Studien verfälschen könnten. Ergänzend wurden qualitative Interviews mit bilingualen Personen durchgeführt, die einige der beobachteten Muster bestätigten. Die Studie zeigt, dass selbst bei formal korrekter Übersetzung bedeutende Unterschiede in der Beantwortung der Fragen auftreten können. Somit konnte auch

gezeigt werden, dass die Validierung von Sprachversionen standardisierter Fragebögen auch DIF-Analysen umfassen sollte.

In der darauffolgenden Studie von 2007 erweiterte der Autor seinen Fokus auf kulturelle und geografische Einflussfaktoren (51). Die Analyse basierte auf Daten aus 106 Studien mit über 28.000 Patienten. Wiederum wurden DIF-Analysen mittels logistischer Regression durchgeführt, diesmal bezogen auf kulturelle bzw. geografische Gruppen. Besonders auffällige Unterschiede in der Beantwortung der Fragen wurden bei Patienten aus Osteuropa und Ostasien festgestellt, während die Antwortmuster aus Großbritannien, den USA und Australien weitgehend übereinstimmten. Interessanterweise ließen sich viele der Unterschiede besser durch die Sprachversion als durch die geografische Herkunft erklären, was die enge Verflechtung von Sprache und Kultur bei der Interpretation von Patientenantworten unterstreicht. Die Studie kommt zu dem Schluss, dass der QLQ-C30 zwar grundsätzlich für internationale Vergleiche geeignet ist, aber kulturell oder sprachlich bedingte Antwortunterschiede berücksichtigt werden müssen, um Verzerrungen in multikulturellen Studien zu vermeiden.

Diese Studien zeigen, dass DIF-Analysen bei der Untersuchung des QLQ-C30 bereits erfolgreich angewendet wurden und ein wichtiges Instrument bei der Fragebogenvalidierung darstellen.

Im Kontext der vorliegenden Studie konnte DIF dazu verwendet werden zu untersuchen, ob sich die Antwortwahrscheinlichkeiten einzelner Items des QLQ-C30 und des QLQ-F17 unterscheiden, obwohl das zugrunde liegende latente Konstrukt, in diesem Fall die Lebensqualität, identisch sind. Sollten sich gerade in den Fragen 8-17 Unterschiede feststellen lassen, so würde dies auf einen Reihenfolgeeffekt hinweisen und die Äquivalenz der beiden Fragebögen in Frage stellen.

Grundlegende Methodik des DIF

Bei der DIF-Analyse unterscheidet man grundsätzlich zwischen uniformer und nicht-uniformer DIF.

Uniforme DIF:

Von uniformer DIF spricht man, wenn sich die Antwortwahrscheinlichkeit eines Items über alle Ausprägungen des latenten Konstrukts hinweg konstant zwischen Gruppen unterscheidet. Das bedeutet, dass die Gruppenunterschiede unabhängig vom Ausprägungsgrad der latenten Fähigkeit (z. B. Lebensqualität) bestehen. Die Differenz in der Beantwortung ist also allein durch die Gruppenzugehörigkeit bedingt.

Nicht-uniforme DIF:

Eine nicht-uniforme DIF liegt vor, wenn sich der Gruppenunterschied in der Antwortwahrscheinlichkeit in Abhängigkeit vom Ausprägungsgrad des latenten Konstrukts verändert. Es besteht somit eine Interaktion zwischen Trait-Level (z. B. individuelle Ausprägung der Lebensqualität) und Gruppenzugehörigkeit.

Die Durchführung der DIF-Analysen erfolgte im Rahmen dieser Auswertung in mehreren aufeinanderfolgenden Schritten und erforderte verschiedene statistische Verfahren. Zur Durchführung der DIF-Analysen wurde das R-Paket *lordif (v0.3–3; Seung W. Choi 2016)* verwendet, welches die einzelnen Schritte durchführt. Im Folgenden werden diese Schritte kurz skizziert.

Ablauf der DIF Analyse

Im ersten Schritt der DIF-Analyse wurde ein IRT-Modell (IRT: Item-Response-Theorie) auf die Gesamtdaten angewendet, um erste Schätzungen der Item-Parameter (Diskriminationsparameter und Schwellenparameter) für die einzelnen Items zu erhalten. Im Rahmen der vorliegenden Studie basierte das IRT-Modell auf dem Graded Response Model (GRM), mit welchem ordinal skalierte Antwortformate adäquat analysiert werden können. Basierend auf diesen ersten Schätzungen wurden für alle Patienten sogenannte Theta-Werte berechnet. Diese Theta-Werte stellen die individuellen Ausprägungen des latenten Merkmals, in diesem Fall der Lebensqualität dar und bildeten die Grundlage für die weiteren Analysen

Im nächsten Schritt wurden für jedes Item drei ordinal logistische Regressionsmodelle unter der Annahme von proportionalen Odds berechnet, welche direkt miteinander verglichen wurden. Zur Überprüfung der Modellannahme von proportionalen Odds wurde der Brant-Test verwendet. Mit diesem Test wird überprüft, ob die Regressionskoeffizienten über alle Antwortkategorien hinweg konstant bleiben. Ein

signifikanter Brant-Test würde auf eine Verletzung der Proportionalitätsannahme hindeuten und damit die Interpretierbarkeit der Modellvergleiche einschränken. In den logistischen Regressionsmodellen wird die kumulative Wahrscheinlichkeit $P(y_i \geq k)$, dass eine Antwort y_i auf Item i in die Kategorie k oder höher fällt modelliert. Die Modelle sind dabei wie folge aufgebaut:

$$\text{Modell 1: } \operatorname{logit}\big(P(y_i \geq k \mid \theta)\big) = \beta_0^{(k)} + \beta_1\theta$$

$$\text{Modell 2: } \operatorname{logit}\big(P(y_i \geq k \mid \theta, G)\big) = \beta_0^{(k)} + \beta_1\theta + \beta_2 G$$

$$\text{Modell 3: } \operatorname{logit}\big(P(y_i \geq k \mid \theta, G)\big) = \beta_0^{(k)} + \beta_1\theta + \beta_2 G + \beta_3(\theta \cdot G)$$

mit:

θ: Latentes Konstrukt (Trait-Level) der Testperson,

G: Gruppenzugehörigkeit (z. B. 0 für Gruppe A, 1 für Gruppe B),

$\beta_0^{(k)}$: Schwellenparameter für Kategorie k,

β_1: Einfluss des latenten Konstrukts,

β_2: Haupteffekt der Gruppe (uniformes DIF),

β_3: Interaktion zwischen latentem Konstrukt und Gruppe (nicht-uniformes DIF).

Im ersten Regressionsmodell wurde die Antwortwahrscheinlichkeit für das jeweilige Item ausschließlich durch den geschätzten Theta-Wert (θ) erklärt. Im zweiten Modell wurde zusätzlich die Gruppenzugehörigkeit (d.h. ob das Item im QLQ-C30 oder im QLQ-F17 beantwortet wurde) als unabhängiger Prädiktor mit aufgenommen. Im dritten Modell wurde darüber hinaus die Interaktion zwischen θ und der Gruppenzugehörigkeit berücksichtigt. Ein signifikanter Unterschied zwischen Modell 1 und Modell 2 wies dabei auf das Vorliegen einer uniformen DIF hin, da sich die Antwortwahrscheinlichkeiten zwischen den Gruppen unabhängig von der Ausprägung des latenten Konstrukts unterschieden. Ein signifikanter Unterschied zwischen Modell 2 und Modell 3 deutete dagegen auf eine nicht-uniforme DIF hin, bei der der Einfluss der Gruppenzugehörigkeit in Abhängigkeit von der Ausprägung des latenten Konstrukts variierte.

Der Vergleich der Modelle erfolgte mittels Likelihood-Ratio-Test (LR-Test), wobei jeweils das einfachere Modell gegen das umfassendere Modell getestet wurde (Modell

1 vs. Modell 2 sowie Modell 2 vs. Modell 3). Als Signifikanzniveau wurde ein konservativer Schwellenwert von $\alpha = 0{,}01$ gewählt. Der LR-Test basierte auf dem Vergleich der modellbezogenen Chi-Quadrat-Werte und identifizierte potenzielle DIF-Items anhand signifikanter Verbesserungen der Modellgüte durch die Aufnahme zusätzlicher Parameter.

Nach der ersten Identifikation potenzieller DIF-Items mittels des Vergleichs der drei logistischen Regressionsmodelle stellte die anschließende iterative Purifikation den nächsten methodischen Schritt dar, um eine stabile und verzerrungsfreie Einschätzung des DIF vorzunehmen. Ziel dieses Verfahrens war es, verzerrungsfreie Schätzungen der latenten Merkmalsausprägung θ zu ermöglichen, indem potenziell DIF-behaftete Items systematisch ausgeschlossen und durch stabile Vergleichsanker ersetzt wurden. Dieses Verfahren begann damit, dass alle Items, bei denen in der ersten Analyse kein signifikanter DIF festgestellt worden war, als vorläufig DIF-frei klassifiziert und als sogenannte Ankeritems definiert wurden. Diese Ankeritems dienten bei der erneuten θ-Schätzung als Referenzbasis, da ihre Itemparameter gruppenunabhängig fixiert blieben. Für die übrigen, als DIF-verdächtig eingestuften Items wurden hingegen gruppenspezifische Parameter geschätzt. Mit den aktualisierten θ-Werten wurde dann eine erneute DIF-Analyse durchgeführt, bei der geprüft wurde, ob sich der Status einzelner Items änderte. Gegebenenfalls wurde die Menge an Ankeritems entsprechend angepasst. Dieser Prozess wiederholte sich so lange, bis sich der Satz an Ankeritems nicht mehr veränderte, also keine neuen DIF-Items hinzukamen oder bestehende entfielen. In diesem Fall galt das Verfahren als konvergiert. Durch diesen iterativen Ansatz wurde die Schätzung der θ-Werte nach und nach verfeinert und die Verzerrung durch falsch klassifizierte Items minimiert.

Ein wesentliches Merkmal der iterativen Purifikation ist, dass die initial als DIF-verdächtig eingestuften Items im Verlauf des Prozesses auch wieder als nicht-DIF-behaftet klassifiziert werden konnten. Dies geschah insbesondere dann, wenn sich herausstellte, dass ihre ursprüngliche Einstufung lediglich auf verzerrten θ-Schätzungen basierte, die durch tatsächlich DIF-behaftete Items beeinflusst worden waren. Durch die sukzessive Verbesserung der θ-Schätzungen auf Basis stabiler Ankeritems wurde die Klassifikation der Items im Verlauf stetig präziser. Somit stellte die iterative Purifikation sicher, dass letztlich nur solche Items als DIF-behaftet

ausgewiesen wurden, bei denen auch unter korrigierten Bedingungen ein signifikanter Gruppenunterschied im Antwortverhalten bestand.

Abschließend erfolgte die Ergebnisinterpretation. Zur Bewertung der Effektstärke potenzieller DIF-Items wurden die zwei Maße Beta-Koeffizient und Pseudo-R^2 nach Nagelkerke herangezogen (52).

1) Relative Veränderung des Beta-Koeffizienten

Dieses Maß gibt an, wie stark sich der Regressionskoeffizient proportional für das latente Merkmal (β_1) ändert, wenn die Gruppenzugehörigkeit (z. B. Fragebogenvariante) zusätzlich als erklärende Variable in das Modell aufgenommen wird. Ein Wert von über 10 % gilt laut gängiger Literatur als Hinweis auf eine potenziell klinisch relevante DIF (53). Niedrigere Werte werden als nicht bedeutsam interpretiert.

2) Pseudo-R^2 nach Nagelkerke

Das Pseudo-R^2 nach Nagelkerke ist über den log-Likelihood der Modelle definiert durch:

$$Pseudo - R^2_{\text{Nagelkerke}} = \frac{1 - \exp\left(-\frac{2}{n}(\text{log-Likelihood}_{\text{Base}} - \text{log-Likelihood}_{\text{Extended}})\right)}{1 - \exp\left(-\frac{2}{n}\text{log-Likelihood}_{\text{Base}}\right)}$$

Dieses Maß ermöglicht eine einfache und intuitive Interpretation der Modellerklärung R^2 des erweiterten Modells im Vergleich zum Basismodell. Es stellt eine skalierte Version des Cox & Snell R^2 dar und nimmt Werte zwischen 0 und 1 an, wobei höhere Werte auf eine stärkere Verbesserung der Modellgüte durch Aufnahme der Gruppenzugehörigkeit bzw. Interaktionseffekte hinweisen. Damit stellt das Pseudo-R^2 nach Nagelkerke (im Weiteren der Einfachheit halber als Pseudo-R^2 bezeichnet) eine praktikable und gut interpretierbare Grundlage zur quantitativen Beurteilung des Ausmaßes von DIF dar. Ein Pseudo-R^2 unter 0,1 wird in der Literatur typischerweise als vernachlässigbar gering eingestuft (54).

Ergänzend wurden grafische Darstellungen, wie Item-True-Score- und Item-Response-Funktionen, eingesetzt, um Unterschiede in den Antwortwahrscheinlichkeiten zwischen den Gruppen anschaulich zu visualisieren.

Multiple lineare Regressionsmodelle

Die multiple lineare Regression ist ein etabliertes Verfahren zur Untersuchung von Zusammenhängen zwischen einer metrischen Zielgröße und mehreren unabhängigen Variablen. In Rahmen dieser Studie diente sie dazu, die Mittelwertunterschiede der Skalenwerte zwischen den beiden Fragebögen (QLQ-C30 vs. QLQ-F17) zu analysieren, wobei für bekannte Störfaktoren kontrolliert wurde. Durch die Adjustierung für relevante Kovariablen konnte gezielt der Effekt der Fragebogenform auf die jeweiligen Skalenwerte isoliert und bewertet werden. Als relevante Kovariablen wurden im Vorfeld Alter, Geschlecht, Land, aktueller Krebsstatus, Q168 („Wie sehr haben Sie unter Nebenwirkungen Ihrer Behandlung gelitten?“), aktuelle Behandlung und Aktivitätsniveau identifiziert.

Grundlage der multiplen linearen Regression ist folgendes Modell:

$$Y_i = \beta_0 + \beta_1 \cdot G_i + \beta_2 \cdot X_{2i} + \beta_3 \cdot X_{3i} + \cdots + \beta_k \cdot X_{ki} + \varepsilon_i$$

mit:

Y_i: Skalenwert des Patienten i (z. B. körperliche Funktion)

β_0: Intercept

β_1: Haupteffekt des Fragebogentyps (geschätzter Mittelwertunterschied der Skala zwischen den beiden Fragebogengruppen unter Kontrolle der übrigen Variablen)

G_i: Gruppenzugehörigkeit (QLQ-C30 = 0, QLQ-F17 = 1)

$\beta_2, \ldots \beta_k$: Effekte der weiteren Kovariablen

$X_{2i}, \ldots, X_{ki}$: Weitere Kovariablen

ε_i: Fehlerterm

Die Modelanpassung erfolgt dabei durch Schätzung der Parameter mittels der Methode der kleinsten Quadrate. Die Regressionskoeffizienten $\beta_1, \ldots \beta_k$ können als erwartete Mittelwertveränderungen des Skalenwerts pro Einheit der jeweiligen

Prädiktoren bei gleichbleibenden Werten der übrigen Variablen interpretiert werden. Bei einer dichotomen Variablen entspricht eine Einheit der Änderung von 0 auf 1.

Zur Beurteilung der Äquivalenz wurde für jede Skala das 95%-Konfidenzintervall des Regressionskoeffizienten β_1 der Gruppenvariable herangezogen. Lag dieses vollständig innerhalb der zuvor definierten Äquivalenzgrenzen von]–5, 5[(siehe 2.4.1 Äquivalenzgrenzen), so wurde die Nullhypothese *H_0: Die Skala ist unterschiedlich zwischen den beiden Fragebogenversionen* verworfen und die Alternativhypothese *H_1: Die Skala ist äquivalent zwischen den beiden Fragebogenversionen* angenommen (siehe Abbildung 6).

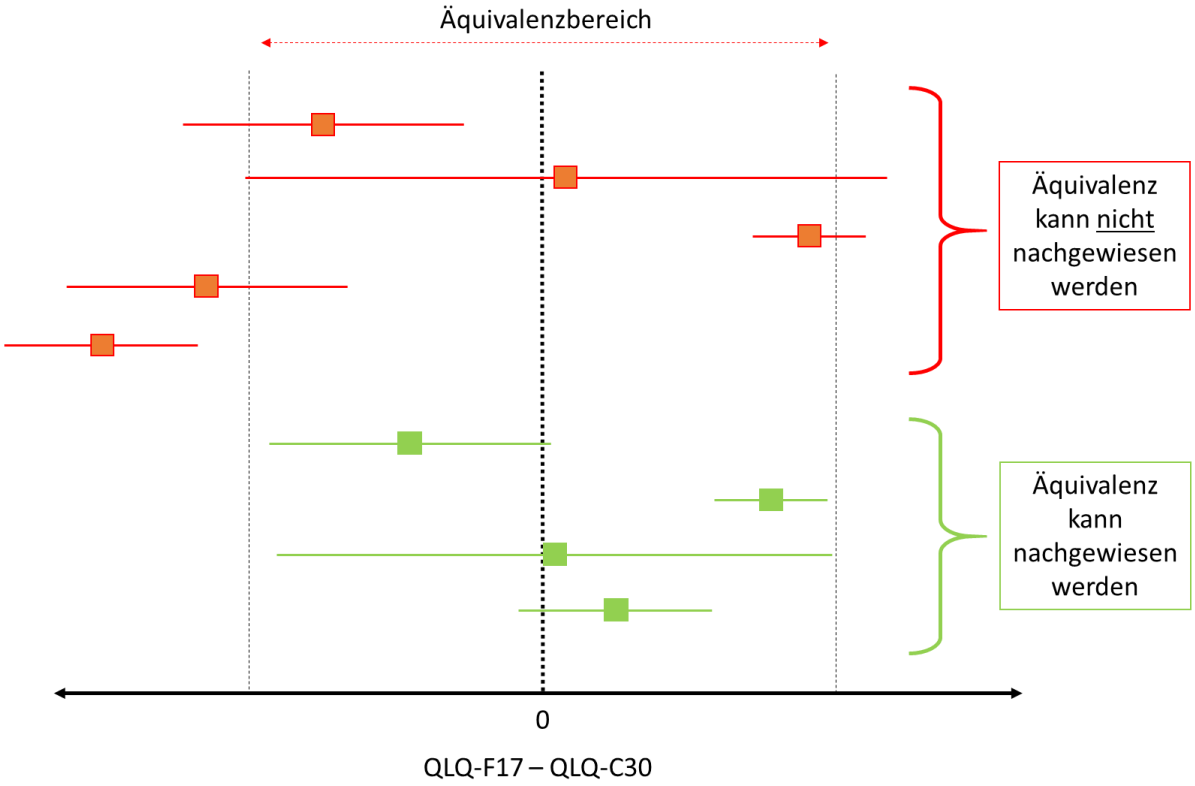


***Abbildung 6**: Beispiele zur Beurteilung von Äquivalenzstudien. Die Quadrate stehen für Effektschätzer (z.B. die mittlere Differenz), die horizontalen Linien für das 95%-Konfidenzintervall. Die senkrechten gestrichelten Linien stehen für den Äquivalenzbereich.*

Bevor jedoch die Ergebnisse der Regressionsmodelle abschließend interpretiert werden konnten, musste überprüft werden, ob die zugrunde liegenden Modellannahmen erfüllt waren. Nur unter diesen Voraussetzungen konnten die geschätzten Effekte, speziell die des Regressionskoeffizienten β_1 der Gruppenvariable

zuverlässig bewerten werden. Lineare Regressionsmodelle basieren dabei auf mehreren statistischen Annahmen, deren Gültigkeit die Verlässlichkeit der Modellschätzungen beeinflusst. Zu den zentralen Voraussetzungen zählen:

1. Linearität zwischen den Prädiktoren und der abhängigen Variable
2. Normalverteilung der Residuen
3. Homoskedastizität, also eine konstante Varianz der Residuen über alle Ausprägungen der Prädiktoren hinweg
4. Keine problematische Multikollinearität zwischen den Prädiktoren

Nur wenn diese Bedingungen zumindest annähernd erfüllt sind, lassen sich die geschätzten Regressionskoeffizienten als valide interpretieren.

Die Annahmen der linearen Regressionsmodelle wurden für alle Modelle überprüft. Hierzu wurden primär diagnostische Plots mithilfe des R-Pakets *performance* erstellt und visuell analysiert (55),(56).

1. Die Linearitätsannahme wurde insbesondere für kontinuierliche Prädiktoren mittels Scatterplots und Component-Residual-Plots überprüft. Diese erlauben eine Einschätzung, ob der Zusammenhang zwischen Prädiktor und abhängiger Variable näherungsweise linear verläuft.
2. Zur Prüfung der Normalverteilungsannahme der Residuen kamen Q-Q-Plots zum Einsatz, in denen die standardnormalverteilten Quantile gegen die beobachteten Quantile der Residuen aufgetragen wurden. Eine annähernd lineare Punktwolke spricht dabei für Normalverteilung.
3. Die Homoskedastizität wurde durch die Analyse der Residuen über den vorhergesagten Werten beurteilt. Da in der Nähe der Grenzen der Skalen (d.h. bei Werten nahe 0 oder 100) Heteroskedastizität zu erwarten war, wurden für die Berechnung der Konfidenzintervalle robuste Standardfehler verwendet.
4. Zur Erkennung potenzieller Multikollinearität wurde für jeden Prädiktor der Varianzinflationsfaktor (VIF) berechnet. Werte oberhalb des Schwellenwerts von 5 wurden als problematische Korrelation zwischen den Prädiktoren definiert.

Eine Übersicht zu den Modellannahmen, den verwendeten Tests und der Bewertung ist in Tabelle 3 dargestellt.

***Tabelle 3**: Übersicht zur Überprüfung und Bewertung der Modellannahmen einer linearen Regression*

Modellannahme	Plot / Test	Bewertung der Annahme
Linearität	Component + Residual Plot, Scatterplot „Residuals vs. Fitted“	Zufällige Streuung der Punkte um eine horizontale Linie spricht für eine lineare Beziehung.
Normalverteilung der Residuen	Quantil-Quantil-Diagramm (Q-Q-Plot)	Liegen die Punkte entlang der Diagonalen ist eine Normalverteilung der Resiuden plausibel.
Homoskedastizität	Residuen vs. vorhergesagte Werte	Ist kein systematisches Muster erkennbar so spricht dies für eine konstante Varianz.
Multikollinearität	Variance Inflation Factor (VIF)	Ist der VIF < 5 liegt keine kritische Multikollinearität vor.

2.6.3.2 Innerhalb-Gruppenvergleiche

Auf Basis der Erhebung beider Fragebögen für jeden Patienten (Innerhalb-Gruppenvergleiche) wurden zur Überprüfung der Äquivalenz zwischen dem QLQ-C30 und dem QLQ-F17 lineare gemischte Modelle berechnet. Zusätzlich wurden weitere Analysen wie die tatsächliche Übereinstimmung auf Item-Ebene, die Test-Retest-Reliabilität auf Item-Ebene und die Reliabilität auf Skalenebene durchgeführt.

Linear gemischte Modelle

Lineare gemischte Modelle (engl. Linear Mixed Models, LMMs) erweitern klassische Regressionsverfahren, indem sie neben festen Effekten auch zufällige Effekte berücksichtigen. Dadurch eignen sie sich besonders für abhängige Datenstrukturen, wie sie bei wiederholten Messungen an denselben Personen vorliegen. In dieser Studie lagen für jede Person zwei Skalenwerte aus zwei unterschiedlichen Fragebogenversionen (QLQ-C30 und QLQ-F17) vor. Aufgrund dieser Messwiederholung sind die Beobachtungen innerhalb einer Person nicht unabhängig, was durch die Einführung zufälliger Effekte für die Subjekte modelliert werden kann.

Folgendes zugrundeliegende Modell wurde für die Analysen verwendet:

$$Y_{ij} = \beta_0 + \beta_1 \cdot F_{ij} + \beta_2 \cdot S_i + \beta_3 \cdot (F_{ij} \times S_i) + u_{0i} + u_{0ij} + \varepsilon_{ij}$$

mit:

Y_{ij}: Skalenwert für Person i in Sequenz j (z. B. körperliche Funktion)

β_0: Intercept

β_1: Haupteffekt des Fragebogentyps (QLQ-C30 vs. QLQ-F17)

β_2: Effekt der Zuordnungssequenz (QLQ-C30 → QLQ-F17 vs. QLQ-F17 → QLQ-C30)

β_3: Interaktionseffekt zwischen Fragebogenversion und Sequenz

u_{0i}: zufälliger Intercept auf Patientenebene

u_{0ij}: zufälliger Intercept auf Sequenzebene innerhalb von Patient i (verschachtelte Struktur)

ε_{ij}: Residualfehler

Der zufällige Intercept erlaubt es, individuelle Unterschiede in den Ausgangswerten der Skalen zu modellieren, wodurch die intraindividuelle Korrelation der Messwerte berücksichtigt wird. Dabei wurde angenommen, dass die zufälligen Intercepts einer Normalverteilung mit dem Mittelwert 0 folgen.

Die verschachtelte Zufallsstruktur ermöglicht es, sowohl individuelle Unterschiede zwischen Patienten als auch Unterschiede innerhalb der Patienten in Abhängigkeit der Zuordnungssequenz zu modellieren. Damit wird die hierarchische Datenstruktur berücksichtigt und die intraindividuelle Korrelation der Messwerte durch getrennte Zufallsintercepts sowie einen Residualfehler adäquat abgebildet.

Ziel der Modelle war es, die Mittelwertdifferenz zwischen den Fragebogenversionen (QLQ-C30 vs. QLQ-F17) unter gleichzeitiger Kontrolle der Zuordnungssequenz zu schätzen. Diese Differenz wurde zusammen mit dem zugehörigen 95 %-Konfidenzintervall interpretiert. Analog zu den zuvor beschriebenen linearen Regressionsmodellen wurde Äquivalenz angenommen, wenn das gesamte

Konfidenzintervall innerhalb der vordefinierten Äquivalenzgrenzen von]–5, 5[Punkten lag.

Analog zu den linearen Regressionsanalysen wurden auch für die linearen gemischten Modelle die zugrundeliegenden statistischen Annahmen geprüft. Diese Annahmen umfassen die Normalverteilung der Residuen, die Varianzhomogenität, die Linearität der Covariablen, sowie die Abwesenheit von Multikollinearität. Dabei wurden die Normalverteilung der Residuen und die Homoskeadastizität analog zu den linearen Regressionsmodellen überprüft. Die Annahmen zur Linearität und Multikollinearität spielten in diesen Modellen keine Rolle, da keine zusätzlichen Covariaten im Modell enthalten waren. Die Analyse erfolgte ebenso anhand diagnostischer Residuenplots, die mit dem R-Paket *performance* erstellt und visuell inspiziert wurden (55),(56).

Bei der Beurteilung der Modellannahmen war zu beachten, dass leichte Verletzungen der Verteilungsannahmen in linearen gemischten Modellen in der Regel nicht zu verzerrten Schätzungen führen (57). Dies unterstreicht die Robustheit dieses Modellansatzes und die Aussagekraft der Ergebnisse.

Übereinstimmung auf Item-Ebene

Zur Beurteilung der Äquivalenz einzelner Fragen wurde die Übereinstimmung der Antworten zwischen dem QLQ-C30 und dem QLQ-F17 auf Item-Ebene berechnet. Für jede Frage wurden zwei Maße der Übereinstimmung ermittelt:

1. der Anteil exakt identischer Antworten sowie
2. der Anteil von Antworten mit einer maximalen Abweichung von einer Antwortkategorie (≤1 Kategorie).

Zur Beurteilung der Anteile wurden die Anteile der ersten sieben Fragen (gleicher Platz und gleiche Reihenfolge in beiden Fragebögen) mit den Anteilen der Fragen 8 bis 17 verglichen.

Test-Retest-Reliabilität auf Item-Ebene

Die weitere Quantifizierung der Konsistenz der Antworten über beide Fragebögen hinweg wurde mittels Test-Retest-Reliabilität erfasst. Hierzu wurde für jedes der 17 Items ein gewichteter Kappa-Koeffizient κ_w berechnet, der Unterschiede in den

Antwortkategorien entsprechend ihrer Distanz gewichtet. Das gewichtete Kappa berücksichtigt sowohl die Häufigkeit der Übereinstimmung als auch die Stärke der Abweichung und bietet damit eine differenzierte Einschätzung der Antwortstabilität:

$$\kappa_w = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} \mathrm{v}_{ij} * f_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} \mathrm{v}_{ij} * e_{ij}}$$

mit:

v_{ij}, quadratische Gewichte

f_{ij}, beobachtete Häufigkeiten

e_{ij}, erwartete Häufigkeiten gemäß Nullhypothese

k, Anzahl der Antwortkategorien

Ein hoher Wert von $\kappa > 0{,}8$ deutet dabei auf eine (fast) perfekte Übereinstimmung hin, während ein Wert zwischen $0{,}61 - 0{,}80$ als eine substantielle Übereinstimmung interpretiert werden kann (58).

Reliabilität auf Skalenebene

Zur Einschätzung der Reliabilität der Skalen wurde der Intraklassen-korrelationskoeffizient (ICC) berechnet. Der ICC quantifiziert den Anteil der Varianz, der auf Unterschiede zwischen den Patienten zurückzuführen ist, im Verhältnis zur Gesamtvarianz. Dabei ist der ICC mathematisch definiert durch:

$$ICC = \frac{\sigma_{zw}^2}{\sigma_{zw}^2 + \sigma_{in}^2}$$

wobei σ_{zw}^2 die Varianz zwischen Patienten und σ_{in}^2 die Varianz innerhalb der Patienten beschreibt.

ICC-Werte unter 0,5 deuten dabei auf eine geringe Reliabilität, Werte zwischen 0,5 und 0,75 auf eine mäßige Reliabilität, Werte zwischen 0,75 und 0,9 auf eine gute Reliabilität und Werte über 0,90 auf eine ausgezeichnete Reliabilität hin (59).

2.6.4 SUBGRUPPENANALYSEN

Zur weiteren Untersuchung der Hauptergebnisse wurden ergänzend Subgruppenanalysen im Rahmen der Zwischen-Gruppenvergleiche durchgeführt. Diese dienten sowohl als Sensitivitätsanalysen zur Prüfung der Robustheit der Befunde als auch der Hypothesengenerierung hinsichtlich möglicher Ursachen der beobachteten DIF-Effekte. Dabei wurden zwei zentrale Subgruppenfaktoren untersucht.

Symptomlast

Der wesentliche Unterschied zwischen den beiden Fragebögen QLQ-F17 und QLQ-C30 liegt im Weglassen der Fragen zu den Symptomen, welche mittig im QLQ-C30 abgefragt werden. Aus den psychologischen Untersuchungen zum Fragebogenverhalten ist bekannt, dass darin ein Einfluss- bzw. Verzerrungspotential auf nachfolgende Fragen bestehen kann (19),(20). Zur Untersuchung eines möglichen Zusammenhangs zwischen Symptomstärke und Ausprägung der DIF-Effekte wurden die Patienten anhand der zusammengefassten Symptomskalenwerte des QLQ-C30 in drei Subgruppen eingeteilt (leichte Symptome, mittlere Symptome, hohe Symptome). Die Einteilung erfolgte auf Basis der Summe der acht symptombezogenen Skalen (Fatigue, Übelkeit/Erbrechen, Schmerzen, Atemnot, Schlaflosigkeit, Appetitlosigkeit, Obstipation, Diarrhoe), wobei das Item zu finanziellen Schwierigkeiten (Frage 28) nicht berücksichtigt wurde. Da jede Skala von 0 bis 100 reicht, ergibt sich ein maximaler Summenwert von 800 Punkten. Patienten mit einer Summe unter 200 wurden der Gruppe mit geringer Symptomlast zugeordnet, Werte zwischen 200 und 400 wurden als mittlere Symptomlast klassifiziert, und Patienten mit Werten über 400 bildeten die Subgruppe mit hoher Symptomlast. Eine weitere Unterteilung innerhalb der Gruppe mit hoher Symptomlast wurde aufgrund der begrenzten Fallzahl nicht vorgenommen. Anschließend wurden DIF-Analysen innerhalb jeder Subgruppe durchgeführt, um zu prüfen, ob sich die Richtung oder Stärke der Item-bezogenen Effekte in Abhängigkeit vom Symptomniveau verändert. Diese Analysen zielen insbesondere darauf ab, die Hypothese zu prüfen, dass eine hohe Symptomlast die Wahrnehmung und Bewertung funktionaler Items beeinflusst.

Geschlecht

Das Geschlecht gilt als bekannter Prädiktor für die gesundheitsbezogene

Lebensqualität. In zahlreichen Studien konnten geschlechtsspezifische Unterschiede in der Wahrnehmung, Bewertung und Berichterstattung von Symptomen und funktionalen Einschränkungen zu Ungunsten von Frauen nachgewiesen werden (60). So tendieren Frauen im Vergleich zu Männern häufig zu geringerer funktionaler Lebensqualität sowie einer höheren Symptombelastung. Entsprechend könnte das Geschlecht auch einen moderierenden Einfluss auf das Antwortverhalten und damit auf potenzielle Reihenfolgeeffekte innerhalb der Fragebögen haben. Basierend darauf wurde analysiert, ob sich die im Hauptergebnis identifizierten Unterschiede im Antwortverhalten zwischen QLQ-F17 und QLQ-C30 konsistent bei Männern und Frauen zeigen. Dazu wurden separate DIF-Analysen für männliche und weibliche Patienten durchgeführt. Ziel war es zu prüfen, ob das Geschlecht als potenzieller Moderator der Reihenfolgeeffekte in Betracht kommt.

Im Gegensatz zu den DIF-Analysen wurden im Rahmen der linearen Regressionsmodelle keine Subgruppenanalysen durchgeführt. Grund hierfür ist die stark eingeschränkte, teils nicht mehr vorhandene statistische Power zum Nachweis von Äquivalenz bei kleinen Subgruppen, da die durch die reduzierten Fallzahlen bedingten Konfidenzintervalle per se breiter ausfallen als der definierte Äquivalenzbereich. Eine valide Bewertung innerhalb der vordefinierten Äquivalenzgrenzen wäre unter diesen Voraussetzungen nicht mehr möglich. Aus methodischen Erwägungen wurden daher keine Subgruppenanalysen in den Regressionsmodellen durchgeführt.

2.6.5 Psychometrische Eigenschaften des QLQ-F17

Für die Analysen der psychometrischen Eigenschaften wurden ausschließlich die Daten aus der ersten Erhebung des QLQ-F17 verwendet. Der Grund hierfür lag in der Vermeidung möglicher Verzerrungen durch die in dieser Arbeit untersuchten Reihenfolgeneffekte sowie möglicher Carry-over-Effekte auf die zweite Erhebung.

Im ersten Schritt wurde eine konfirmatorische Faktorenanalyse (CFA) durchgeführt. Hierfür wurde ein sechsfaktorielles Modell mit korrelierten Faktoren spezifiziert, entsprechend der sechs Skalen des QLQ-F17 und analog zur Struktur des QLQ-C30 (siehe *2.3.1 EORTC QLQ-C30 (Version 3.0)*). Zur Bewertung der Modellgüte der CFA wurden zum einen die standardisierten Faktorladungen sowie die folgenden Fit-Indizes

berechnet: der Comparative Fit Index (CFI), der Tucker-Lewis Index (TLI), der Root Mean Square Error of Approximation (RMSEA) sowie das Standardized Root Mean Square Residual (SRMR). Die Analysen erfolgten mit dem R-Paket *lavaan* (Version 0.6-17).

Für die Interpretation der Ergebnisse wurden die Grenzwerte nach Kline und Byrne herangezogen (61),(62). Diese definieren die Grenzen für einen akzeptablen Model-Fit durch einen CFI ≥ 0,85, einen TLI ≥ 0,85, eine RMSEA < 0,08 und einen SRMR < 0,08. Die standardisierten Faktorladungen geben die Korrelation zwischen Items und dem jeweiligen Faktor (Skala) wieder. Eine Ladung ≥ 0,40 wurde als hinreichender Nachweis für eine ausreichende Item-Faktor-Zuordnung gewertet (63).

Die Reliabilität (interne Konsistenz) der Skalen des QLQ-F17 wurde mittels Cronbachs Alpha bestimmt. Gemäß Konvention gelten Alpha-Werte ≥ 0,70 als Indikator für eine akzeptable interne Konsistenz (64).

Zur Prüfung der Konstruktvalidität wurde eine Multi-Trait-Skalierungsanalyse durchgeführt. Die konvergente Validität eines Items war gegeben, wenn die korrigierte Korrelation mit der zugehörigen Skala > 0,40 betrug (63). Die korrigierte Korrelation bedeutet dabei, dass das jeweilige Item beim Berechnen des Skalenwertes nicht berücksichtigt wurde. Die diskriminante Validität wurde durch den Vergleich der Item-Skalen-Korrelation mit den Korrelationen zu anderen Skalen geprüft. Ein Skalierungsfehler lag vor, wenn ein Item stärker mit einer fremden Skala als mit seiner eigenen korrelierte.

3 ERGEBNISSE

3.1 DEFINITION DER ANALYSEPOPULATION

Im Zeitraum vom 17.02.2023 bis zum 28.03.2023 haben insgesamt 5.668 Patienten auf die Umfrage zugegriffen und gestartet. Dabei konnten 2.810 (50%) Patienten bei der Screeningfrage keine Krebsdiagnose bestätigen, so dass diese als Screeningfehler direkt ausgeschlossen wurden. Von den 2.858 Patienten, welche die Umfrage erfolgreich gestartet haben, haben 186 (6,5%) Patienten die Studie vorzeitig abgebrochen, so dass insgesamt 2.672 Patienten erfolgreich an der Studie teilgenommen und beide Fragebögen ausgefüllt haben. Die Kennzahlen zur Patientenrekrutierung sind in Tabelle 4 aufgeteilt nach jeweiligem Land dargestellt.

***Tabelle 4**: Kennzahlen zur Patientenrekrutierung*

Land	Ziel-population	Anzahl Zugriffe	Umfrage beendet	Abbrecher	Screening-fehler	Teilnahme-rate
Australien	150	234	150	7	77	64%
Deutschland	300	368	300	7	61	82%
Finnland	150	602	150	15	437	25%
Frankreich	300	368	301	16	51	82%
Italien	300	387	300	6	81	78%
Polen	250	1.270	250	16	1.004	20%
Rumänien	150	399	150	18	231	38%
Schweden	150	394	150	17	227	38%
Spanien	300	701	300	26	375	43%
UK	200	238	200	12	26	84%
USA	420	707	421	46	240	60%
Gesamt	**2.670**	**5.668**	**2.672**	**186**	**2.810**	**47%**

Im Anschluss an die Datenerhebung erfolgte eine systematische Qualitätskontrolle der erhobenen Daten (siehe Kapitel *2.4.2 Analysepopulation*), um die finale Analysepopulation zu definieren.

Im ersten Schritt der Qualitätsprüfung wurde für jede der ersten sieben Fragen des QLQ-C30 und QLQ-F17 überprüft, ob gegensätzliche Antworten vorlagen. Patienten, bei denen dies zutraf, erhielten für die jeweilige Frage eine Markierung (siehe Abbildung 7). Je nach Frage waren zwischen 4 und 13 Patienten betroffen (siehe Tabelle 5).

Tabelle 5: *Qualitätsüberprüfung der Daten basierend auf gegensätzlichen Antworten bei den ersten 7 Fragen (1)*

Frage QLQ-C30/ QLQ-F17	Antworten im Vergleich QLQ-C30 vs QLQ-F17	Anzahl markierter Patienten
#1: Bereitet es Ihnen Schwierigkeiten, sich körperlich anzustrengen (z. B. eine schwere Einkaufstasche oder einen Koffer zu tragen)?	Sehr vs. überhaupt nicht Überhaupt nicht vs. sehr	5 (0,2%) 2 (0,1%)
#2: Bereitet es Ihnen Schwierigkeiten, einen längeren Spaziergang zu machen?	Sehr vs. überhaupt nicht Überhaupt nicht vs. sehr	4 (0,2%) 0 (0,0%)
#3: Bereitet es Ihnen Schwierigkeiten, eine kurze Strecke außer Haus zu gehen?	Sehr vs. überhaupt nicht Überhaupt nicht vs. sehr	5 (0,2%) 4 (0,2%)
#4: Müssen Sie tagsüber im Bett liegen oder in einem Sessel sitzen?	Sehr vs. überhaupt nicht Überhaupt nicht vs. sehr	4 (0,2%) 0 (0,0%)
#5: Brauchen Sie Hilfe beim Essen, Anziehen, Waschen oder Benutzen der Toilette?	Sehr vs. überhaupt nicht Überhaupt nicht vs. sehr	5 (0,2%) 2 (0,1%)
#6: Waren Sie bei Ihrer Arbeit oder bei anderen tagtäglichen Beschäftigungen eingeschränkt?	Sehr vs. überhaupt nicht Überhaupt nicht vs. sehr	3 (0,1%) 10 (0,4%)
#7: Waren Sie bei Ihren Hobbys oder anderen Freizeitbeschäftigungen eingeschränkt?	Sehr vs. überhaupt nicht Überhaupt nicht vs. sehr	8 (0,3%) 5 (0,2%)

Zusätzlich wurden die Antwortzeiten analysiert. Hierfür wurden sowohl die Gesamtzeit der Befragung als auch die durchschnittliche Antwortzeit pro Frage und LQ-Fragebogen berechnet. Bei einer Medianzeit von 11:47 Minuten wurden Patienten mit einer Gesamtzeit von weniger als 5:54 Minuten markiert, was auf eine potenziell unzureichende Auseinandersetzung mit den Inhalten hinweist. Dieses Kriterium traf auf 146 Patienten (5,5 %) zu. Die Verteilung der Antwortzeiten ist in Abbildung 8 dargestellt.

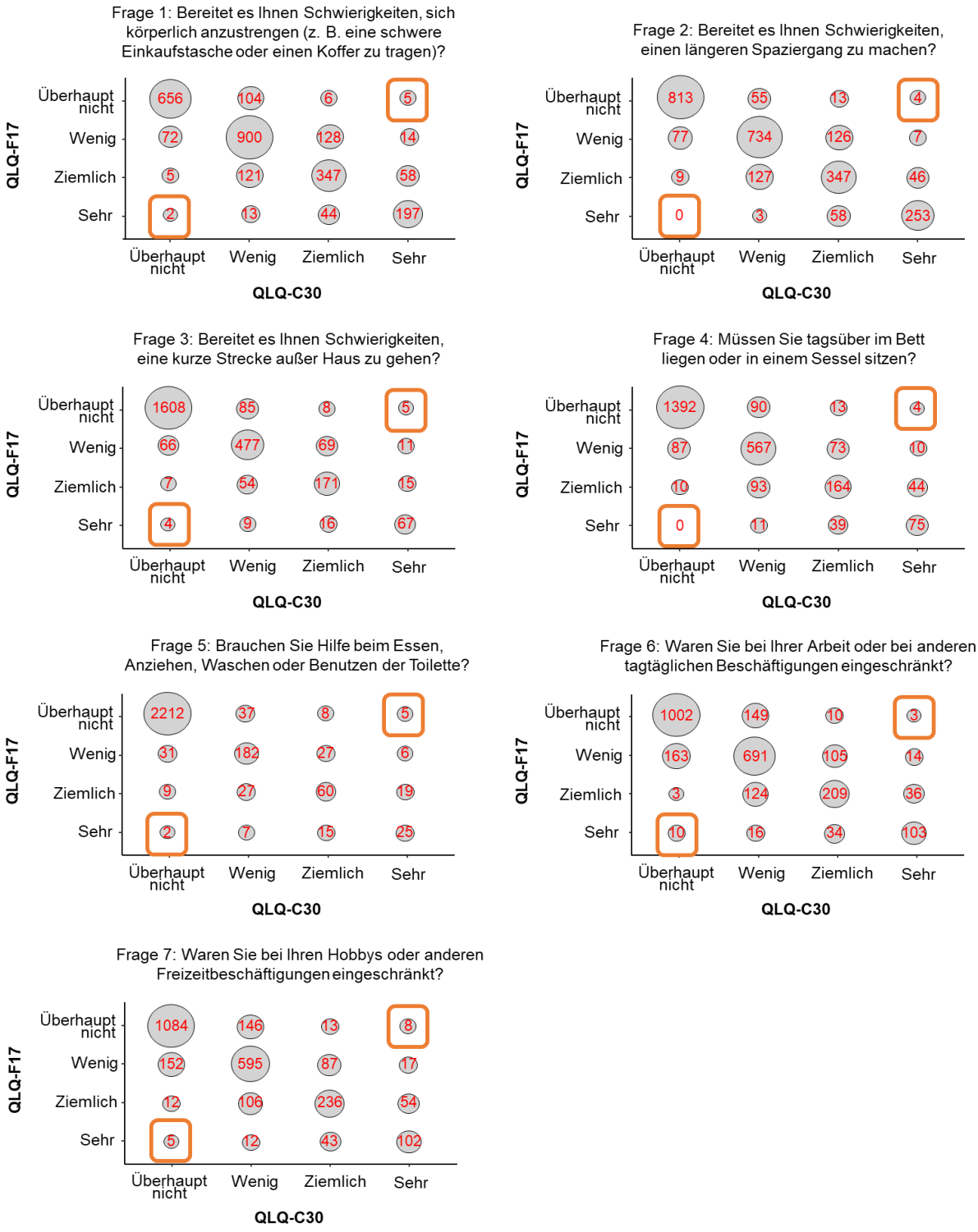


***Abbildung 7**: Vergleich der Antworten der ersten 7 Items des QLQ-F17 und des QLQ-C30*

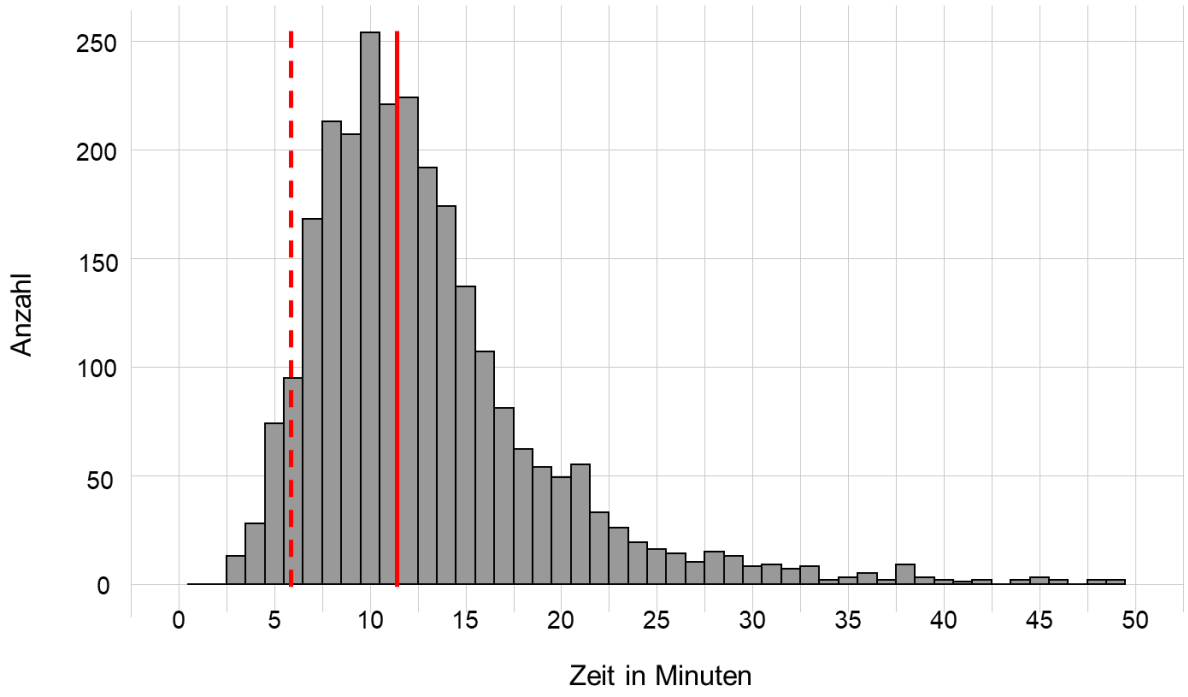


***Abbildung 8**: Histogramm zur Gesamtbearbeitungszeit der Befragung. Die durchgezogene rote Linie entspricht dem Median von 11:47 Minuten. Die gestrichelte rote Linie dem halben Median von 5:54 Minuten.*

Bei der mittleren Antwortzeit je Frage galt ein Schwellenwert von unter 2 Sekunden als kritisch. Dies entspricht einer Gesamtbearbeitungszeit von unter einer Minute für den QLQ-C30 (siehe Abbildung 9) sowie unter 34 Sekunden für den QLQ-F17 (siehe Abbildung 10). Entsprechend wurden 69 Patienten (2,6 %) beim QLQ-C30 und 62 Patienten (2,3 %) beim QLQ-F17 markiert (Tabelle 6).

***Tabelle 6**: Qualitätsüberprüfung der Daten basierend auf den Antwortzeiten (1)*

Zeitmessung	Cut-off	Anzahl markierter Patienten
Gesamtdauer der Session	<5:54 Minuten*	146 (5,5%)
Durchschnittliche Antwortzeit QLQ-C30	<2 Sekunden	69 (2,6%)
Durchschnittliche Antwortzeit QLQ-F17	<2 Sekunden	62 (2,3%)

** Hälfte der medianen Zeit der gesamten Stichprobe*

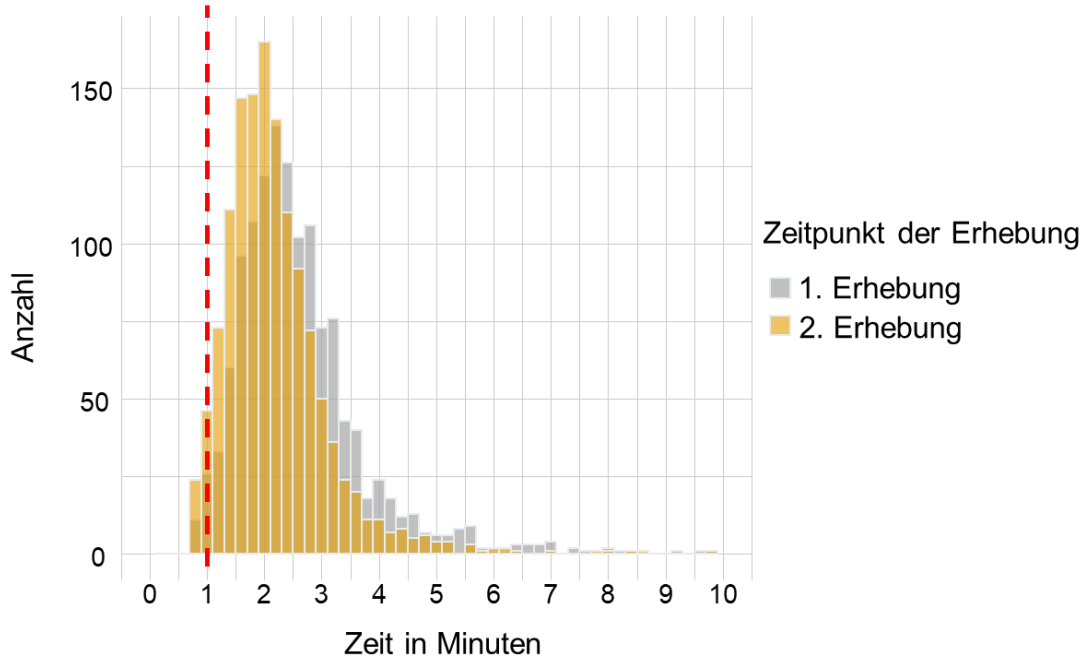


***Abbildung 9**: Histogramm zur Bearbeitungszeit des QLQ-C30 unterteilt nach Zeitpunkt der Erhebung. Die rote gestrichelte Linie kennzeichnet die Gesamtbeantwortungszeit von einer Minute, was einer durchschnittlichen Zeit pro Frage von 2 Sekunden entspricht.*

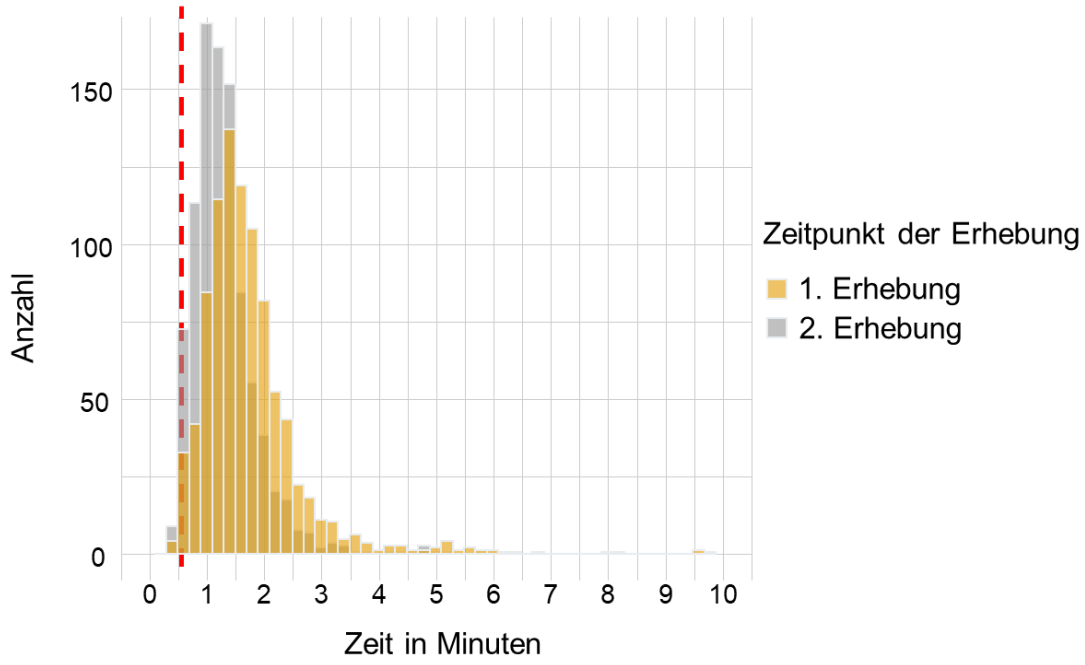


***Abbildung 10**: Histogramm zur Bearbeitungszeit des QLQ-F17 unterteilt nach Zeitpunkt der Erhebung. Die rote gestrichelte Linie kennzeichnet die Gesamtbeantwortungszeit von 34 Sekunden, was einer durchschnittlichen Zeit pro Frage von 2 Sekunden entspricht.*

Gemäß der vordefinierten Qualitätskriterien wurden alle Patienten mit drei oder mehr Markierungen von der Analyse ausgeschlossen. Insgesamt betraf dies n = 29 Patienten. Die finale Analysepopulation umfasste somit n = 2.643 Patienten (siehe Tabelle 7).

***Tabelle 7**: Definition der Analysepopulation basierend auf den Datenqualitätsüberprüfungen (1)*

Gesamtzahl an Markierungen	Anzahl Patienten (%)	Entscheidung
0	2469 (92,4%)	Analysepopulation (n=2.643)
1	113 (4,2%)	
2	61 (2,3%)	
3	22 (0,8%)	Von der Analysepopulation ausgeschlossen (n=29)
4	4 (0,2%)	
5	2 (0,1%)	
6	0 (0,0%)	
7	1 (0,04%)	
8	0 (0,0%)	
9	0 (0,0%)	
10	0 (0,0%)	

Der vollständige Patientenfluss von Studienanfrage bis zur finalen Analysepopulation ist nach Vorgabe der Erweiterung des CONSORT Statements für randomisierte Cross-over Studien in Abbildung 11 dargestellt (65).

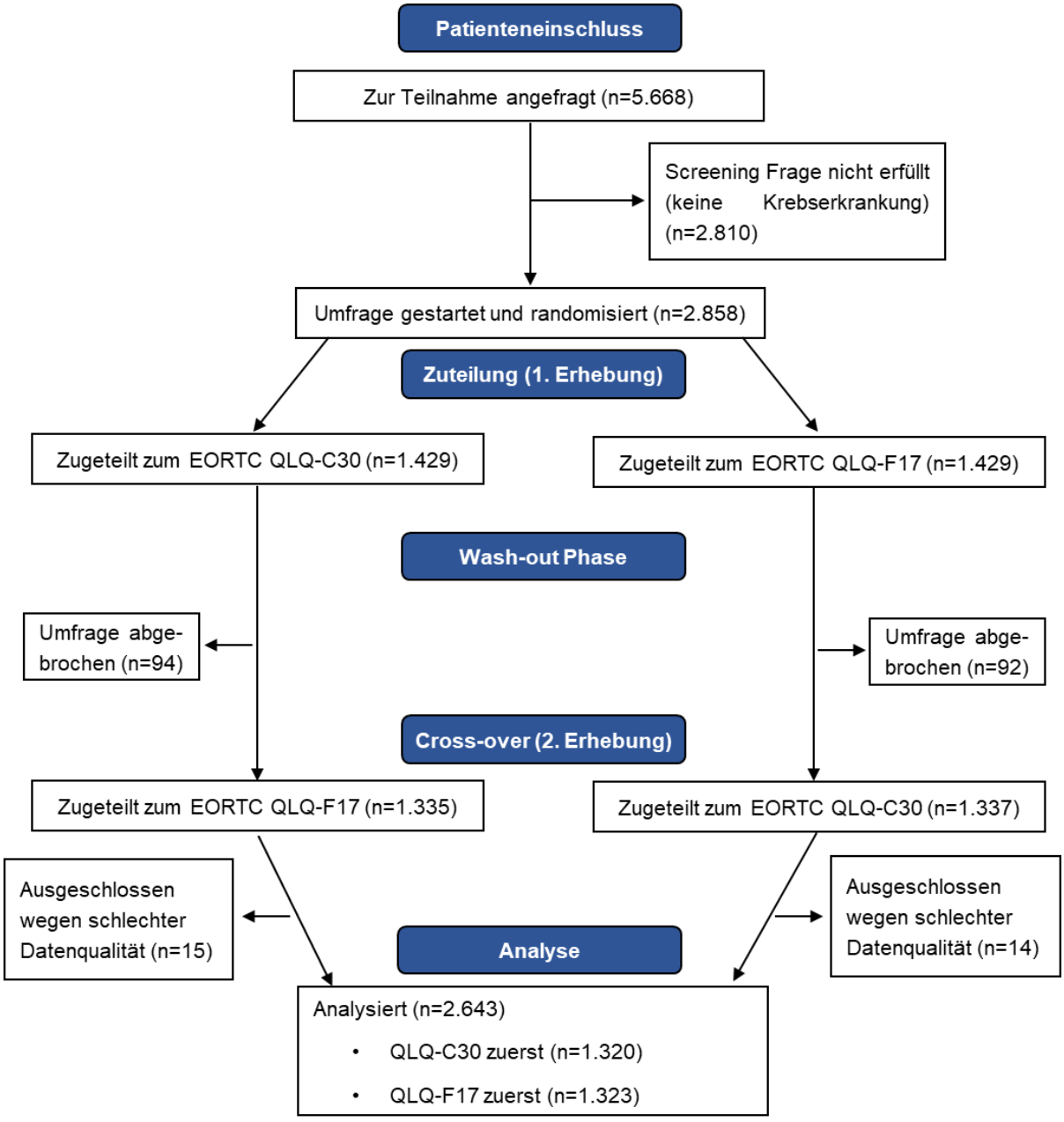


***Abbildung 11**: Flussdiagramm des Patientenverlaufs*

3.2 BESCHREIBUNG DER ANALYSEPOPULATION

Die finale Stichprobe umfasste Patienten aus 11 Ländern und 6 verschiedenen Sprachregionen. Das Durchschnittsalter betrug 58 Jahre mit einer Spanne von 18 bis 92 Jahre. Die Geschlechterverteilung war ausgeglichen, mit jeweils 50 % männlichen und weiblichen Patienten.

Zum Zeitpunkt der Befragung befanden sich 18 % der Patienten in aktiver Krebsbehandlung, 74 % waren in Remission, und 7 % gaben an, innerhalb der letzten drei Monate eine neue Krebsdiagnose erhalten zu haben. Die Stichprobe umfasste alle Hauptkrebsentitäten, wobei Brustkrebs (21 %) und Prostatakrebs (14 %) am häufigsten vertreten waren.

Eine detaillierte Übersicht über soziodemografische und klinische Merkmale der Patienten findet sich in der folgenden Tabelle 8.

***Tabelle 8**: Patienten Charakteristiken (n=2.643) (1)*

	Gesamte Stichprobe (n=2.643)	QLQ-C30 zuerst (n=1.320)	QLQ-F17 zuerst (n=1.323)
Geschlecht, n (%)			
Männlich	1.311 (50%)	650 (49%)	661 (50%)
Weiblich	1.323 (50%)	664 (50%)	659 (50%)
Ich identifiziere mich nicht als eines von beiden	7 (0,2%)	4 (0,3%)	3 (0,2%)
Keine Angabe	2 (0,1%)	2 (0,1%)	0 (0%)
Alter, Mittelwert ± SD (min, max)	58 ±15 (18, 92)	58 ±16 (18, 90)	58 ±15 (18, 92)
Krebsstatus, n (%)			
Bei mir wurde in den letzten 3 Monaten zum ersten Mal Krebs diagnostiziert.	190 (7%)	103 (8%)	87 (7%)
Ich befinde mich derzeit in Krebstherapie.	488 (18%)	257 (19%)	231 (17%)
Ich bin in Remission / Ich habe meinen Krebs besiegt.	1.965 (74%)	960 (73%)	1.005 (76%)

In welchem Ausmaß fühlten Sie sich durch Nebenwirkungen Ihrer Behandlung belästigt? (Q168)?, n (%)			
Überhaupt nicht	932 (35%)	480 (36%)	452 (34%)
Wenig	988 (37%)	493 (37%)	495 (37%)
Ziemlich	539 (20%)	269 (20%)	270 (20%)
Sehr	184 (7%)	78 (6%)	106 (8%)
Krebsart, n (%)			
Brust	564 (21%)	284 (22%)	280 (21%)
Prostata	373 (14%)	180 (14%)	193 (15%)
Haut	268 (10%)	126 (10%)	142 (11%)
Andere gynäkologische Tumoren (z. B. Gebärmutter, Eierstöcke)	259 (10%)	125 (10%)	134 (10%)
Darm	208 (8%)	104 (8%)	104 (8%)
Lunge	181 (7%)	103 (8%)	78 (6%)
Blase	79 (3%)	46 (4%)	33 (3%)
Niere	76 (3%)	32 (2%)	44 (3%)
Leukämie	71 (3%)	35 (3%)	36 (3%)
Magen	70 (3%)	33 (3%)	37 (3%)
Lymphom	69 (3%)	36 (3%)	33 (3%)
Schilddrüse	68 (3%)	30 (2%)	38 (3%)
Speiseröhre	64 (2%)	33 (3%)	31 (2%)
Hoden	41 (2%)	20 (2%)	21 (2%)
Leber	39 (2%)	21 (2%)	18 (1%)
Gehirn	30 (1%)	16 (1%)	14 (1%)
Myelom	26 (1%)	13 (1%)	13 (1%)
Bauchspeicheldrüse	17 (1%)	9 (1%)	8 (1%)
Andere Indikation*	140 (5%)	74 (6%)	66 (5%)
Zeitpunkt der Krebsdiagnose, n (%)			
Innerhalb der letzten 6 Monate	204 (8%)	104 (8%)	100 (8%)
Innerhalb des letzten Jahres	278 (11%)	141 (11%)	137 (10%)
Innerhalb der letzten 2 Jahre	440 (17%)	220 (17%)	220 (17%)
Innerhalb der letzten 5 Jahre	642 (24%)	319 (24%)	323 (24%)
Innerhalb der letzten 10 Jahre	493 (19%)	243 (18%)	250 (19%)

Vor mehr als 10 Jahren	586 (22%)	293 (22%)	293 (22%)
Art der aktiven Behandlung (mehrere Antworten möglich), n (%)			
Chemotherapie	291 (11%)	150 (11%)	141 (11%)
Immun- oder zielgerichtete Therapie	283 (11%)	135 (10%)	148 (11%)
Radiotherapie	184 (7%)	89 (7%)	95 (7%)
Operation innerhalb der letzten 3 Monate	157 (6%)	81 (6%)	76 (6%)
Aktivitätsgrad, n (%)			
Voll aktiv und fähig, alle Tätigkeiten ohne Einschränkung durchzuführen	607 (23%)	303 (23%)	304 (23%)
Aktiv, mit leichter Einschränkung bei körperlich anstrengenden Tätigkeiten	1.020 (39%)	504 (38%)	516 (39%)
Einschränkungen in der Aktivität und bei körperlich anstrengenden Tätigkeiten	698 (26%)	348 (26%)	350 (26%)
Selbstversorgung möglich, aber nicht fähig, einer Arbeitstätigkeit nachzugehen	249 (9%)	127 (10%)	122 (9%)
Eingeschränkte Selbstversorgung mit mehr als 50% des Tages im Bett oder im Stuhl	69 (3%)	38 (3%)	31 (2%)
Höchster abgeschlossener Bildungsstand, n (%)			
Real- oder Mittelschule	367 (14%)	189 (14%)	178 (13%)
Gymnasium	898 (34%)	461 (35%)	437 (33%)
Universitätsabschluss	1.115 (42%)	548 (42%)	567 (43%)
Promotion	90 (3%)	40 (3%)	50 (4%)
Anderer	161 (6%)	76 (6%)	85 (6%)
Keine Angabe	12 (0,5%)	6 (0,5%)	6 (0,5%)
Land, n (%)			
Australien	150 (6%)	75 (6%)	75 (6%)
Deutschland	297 (11%)	149 (11%)	148 (11%)
Finnland	147 (6%)	75 (6%)	72 (5%)
Frankreich	298 (11%)	149 (11%)	149 (11%)
Italien	297 (11%)	148 (11%)	149 (11%)
Polen	244 (9%)	121 (9%)	123 (9%)
Rumänien	147 (6%)	73 (6%)	74 (6%)
Schweden	147 (6%)	73 (6%)	74 (6%)

Spanien	297 (11%)	148 (11%)	149 (11%)
Vereinigtes Königreich	199 (8%)	99 (8%)	100 (8%)
Vereinigte Staaten von Amerika	420 (16%)	210 (16%)	210 (16%)
Sprachregionen, n (%)			
Romanische Sprachen: Französisch (Frankreich), Italienisch, Spanisch (Spanien), Rumänisch	1039 (39%)	518 (39%)	521 (39%)
Englischsprachige Länder: UK, USA, Australien	769 (29%)	384 (29%)	385 (29%)
Westgermanische Sprache: Deutsch	297 (11%)	149 (11%)	148 (11%)
Slawische Sprachen: Polnisch, Russisch, Ukrainisch	244 (9%)	121 (9%)	123 (9%)
Skandinavische Sprache: Schwedisch	147 (6%)	73 (6%)	74 (6%)
Andere europäische Sprache: Finnisch	147 (6%)	75 (6%)	72 (5%)

**andere wurden als Freitext angegeben und umfassen seltene Krebserkrankungen wie Zungenkrebs, Plattenepithelkarzinom oder Gallenblasenkrebs oder waren aufgrund unverständlicher Antworten nicht kategorisierbar.*

Die mediane Zeit für die gesamte Umfrage betrug 11:49 Minuten [IQR: 8:56 – 15:35]. Die mittleren Beantwortungszeiten für jede Frage in den Fragebögen, für die Zwischenfragen (Wash-out Phase) und für die gesamte Umfrage sind in Tabelle 9 aufgeführt.

***Tabelle 9**: Dauer der Erhebung der einzelnen Fragebögen, der einzelnen Fragen und der gesamten Umfrage (1)*

		10th Perzentil in mm:ss	Median in mm:ss	90th Perzentil in mm:ss
EORTC QLQ-C30	Insgesamt	01:24 (00:03)	02:14 (00:04)	03:39 (00:07)
	Erste Erhebung	01:33 (00:03)	02:26 (00:05)	04:06 (00:08)
	Zweite Erhebung	01:19 (00:03)	02:03 (00:04)	03:15 (00:06)
EORTC QLQ-F17	Insgesamt	00:49 (00:03)	01:24 (00:05)	02:23 (00:08)
	Erste Erhebung	00:59 (00:03)	01:36 (00:06)	02:44 (00:10)
	Zweite Erhebung	00:45 (00:03)	01:13 (00:04)	01:59 (00:07)

Zeit für die Wash-out Phase	04:21	07:36	14:54
Gesamtzeit der Umfrage	07:03	11:49	21:29

Daten zeigen Gesamtzeit (Zeit pro item), Die Zeit ist dargestellt als Minuten:Sekunden (mm:ss)

3.3 ÄQUIVALENZ - ZWISCHEN-GRUPPENVERGLEICHE

3.3.1 DIFFERENZIELLE ITEM-FUNKTION

Zur Untersuchung der Äquivalenz auf Item-Ebene wurde analysiert, ob bestimmte Fragen des QLQ-F17 unabhängig von der zugrunde liegenden Merkmalsausprägung der Patienten systematisch andere Antwortmuster hervorrufen als die entsprechenden Fragen des QLQ-C30. Ein solcher Unterschied weist darauf hin, dass sich die Interpretation oder das Antwortverhalten auf eine Frage in Abhängigkeit vom verwendeten Fragebogen unterscheiden könnte. Unterschieden wird dabei zwischen uniformer DIF, bei der der Effekt unabhängig vom Ausprägungsgrad des latenten Konstrukts ist, und nicht-uniformer DIF, bei der sich die Effekte je nach Merkmalsausprägung verändern. Die Identifikation potenzieller DIF war entscheidend, um sicherzustellen, dass der QLQ-F17 die gleichen Inhalte misst wie der etablierte QLQ-C30.

Für die Identifikation der DIF-Items wurden drei iterative Analyse-Schritte durchgeführt. Während im ersten Schritt noch sechs Items als DIF-verdächtig identifiziert wurden, verblieben nach dem dritten Schritt insgesamt fünf Items mit statistisch signifikanter DIF. Folgende fünf Items zeigten dabei eine uniforme DIF:

- Item Nr. 8 („Hatten Sie Schwierigkeiten, sich zu konzentrieren, z. B. beim Lesen einer Zeitung oder beim Fernsehen?“),
- Item Nr. 9 („Fühlten Sie sich angespannt?“),
- Item Nr. 10 („Haben Sie sich Sorgen gemacht?“),
- Item Nr. 14 („Hat Ihr körperlicher Zustand oder Ihre medizinische Behandlung Ihr Familienleben beeinträchtigt?“) sowie
- Item Nr. 15 („Hat Ihr körperlicher Zustand oder Ihre medizinische Behandlung Ihre sozialen Aktivitäten beeinträchtigt?“).

Darüber hinaus wurde für Item Nr. 14 zusätzlich eine nicht-uniforme DIF identifiziert, d. h. die Richtung der Unterschiede zwischen den Fragebögen variierte je nach Ausprägung des zugrunde liegenden Merkmals (siehe Tabelle 10).

***Tabelle 10**: Ergebnisse der DIF-Analyse (1)*

Item # F17	Item # C30	Skala	%Beta	Effekt* uniform	p-Wert uniform	Effekt* nicht-uniform	p-Wert nicht-uniform
1	1	PF	<0,001	<0,001	0,60	<0,001	0,44
2	2	PF	<0,001	<0,001	0,30	<0,001	0,37
3	3	PF	<0,001	<0,001	0,38	<0,001	0,41
4	4	PF	<0,001	<0,001	0,65	<0,001	0,93
5	5	PF	<0,001	<0,001	0,66	0,002	0,026
6	6	RF	<0,001	<0,001	0,93	<0,001	0,87
7	7	RF	<0,001	<0,001	0,76	<0,001	0,49
8	20	CF	0,012	0,008	**<0,001**	<0,001	0,20
9	21	EF	0,008	0,005	**<0,001**	<0,001	0,83
10	22	EF	0,008	0,007	**<0,001**	<0,001	0,78
11	23	EF	<0,001	<0,001	0,50	<0,001	0,78
12	24	EF	<0,001	<0,001	0,19	<0,001	0,58
13	25	CF	0,002	0,001	0,06	<0,001	0,74
14	26	SF	0,010	0,005	**<0,001**	0,001	**0,008**
15	27	SF	0,008	0,003	**<0,001**	<0,001	0,55
16	19	QL	0,001	0,001	0,090	<0,001	0,80
17	30	QL	<0,001	<0,001	0,45	<0,001	0,83

*Pseudo-R^2 *nach Nagelkerke; %Beta, relative Veränderung des Beta-Wertes von Modell 1 zu Modell 2; PF, Physische Funktion; RF, Rollenfunktion; EF, emotionale Funktion; CF, kognitive Funktion; SF, soziale Funktion; QL, globaler Gesundheitszustand/ Lebensqualität. Die Ergebnisse beziehen sich auf die erste Erhebung und vergleichen die Patienten, die entweder den EORTC QLQ-C30 (n = 1.320) oder den EORTC QLQ-F17 (n = 1.323) ausgefüllt haben. Graue Zeilen sind die Fragen, die in beiden Fragebögen an derselben Position stehen. Rote Zeilen unterscheiden sich in der Position zwischen QLQ-C30 und QLQ-F17. Ein p-Wert < 0,01 weist auf eine uniforme oder nicht-uniforme DIF einer Frage hin. Eine Effektgröße < 0,01 kann als trivial und nicht aussagekräftig angesehen werden.*

Die proportionale Veränderung des Beta Koeffizienten war bei allen fünf identifizierten Items sehr gering, mit einer maximalen Änderung von 0,012 (1,2 %) für Item Nr. 8, das direkt im Anschluss an die Symptomskala des QLQ-C30 positioniert ist. Somit deuten die beobachteten Veränderungen auf keine bedeutsamen Unterschiede zwischen den Fragebögen hin. Auch die Pseudo-R^2 Werte lagen bei allen Items unterhalb von 0,01, was für eine sehr geringe Ausprägung der DIF spricht.

Zur visuellen Beurteilung von DIF wurden für alle identifizierten Items jeweils zwei diagnostische Diagramme erstellt (Abbildung 12 - Abbildung 16). Im oberen Diagramm, der sogenannten „Item True Score Funktionen“, zeigen die schwarze und die rote Linie die Wahrscheinlichkeitskurven für die Beantwortung der entsprechenden Items im QLQ-C30 und im QLQ-F17 in Abhängigkeit vom Ausprägungsgrad des latenten Konstrukts (θ). Sind die Kurven parallel zueinander verschoben, deutet dies auf eine uniforme DIF hin.

Betrachtet man beispielsweise Abbildung 12 zur Untersuchung von Item 8 im QLQ-F17 (F8) im Vergleich zu Item 20 im QLQ-C30 (C20), zeigt sich, dass die Kurve des QLQ-F17 gegenüber jener des QLQ-C30 parallel nach rechts verschoben ist. Dies weist darauf hin, dass die Beantwortung von Frage 8 im QLQ-F17 tendenziell besser ausfällt als jene von Frage 20 im QLQ-C30 unabhängig vom Ausprägungsgrad des latenten Konstrukts (uniforme DIF)**.**

In Abbildung 14 zur Untersuchung von Item 14 im QLQ-F17 (F14) im Vergleich zu Item 26 im QLQ-C30 (C26) hingegen kreuzen sich die Kurven. Dies spricht für eine nicht-uniforme DIF**.** Bei niedrigen θ-Werten fällt die Antwort auf Frage F14 im Vergleich zu C26 schlechter aus, während sich die Unterschiede mit steigendem θ annähern und schließlich sogar leicht umkehren.

Im unteren Diagramm, der „Item Response Funktionen“, werden die Wahrscheinlichkeiten für jede der vier Antwortoptionen („überhaupt nicht“, „wenig“, „ziemlich“, „sehr“) in Abhängigkeit vom latenten Merkmal dargestellt. Eine parallele Verschiebung dieser Kurven wie in Abbildung 12 zeigt, dass für ein bestimmtes θ die Wahrscheinlichkeit, eine höhere Antwortkategorie zu wählen, im QLQ-F17 tendenziell größer ist als im QLQ-C30. Bei nicht-uniformer DIF, wie in Abbildung 14, erkennt man, wie sich diese Wahrscheinlichkeiten mit steigendem θ angleichen oder verschieben.

Betrachtet man die diagnostischen Diagramme, so zeigt sich eine differenzierte Ausprägung der DIF n Abhängigkeit von der jeweiligen Skala:

- Item Nr. 8 (kognitive Funktion) des QLQ-F17 weist eine leichte uniforme DIF zugunsten des QLQ-F17 auf, d. h. die Werte bei den entsprechenden Items des QLQ-C30 lagen systematisch etwas höher.
- Bei Item Nr. 9 und Nr. 10 (beide emotionale Funktion) und Item Nr. 15 (soziale Funktion) des QLQ-F17 zeigte sich eine leichte uniforme DIF zugunsten des QLQ-C30, mit tendenziell niedrigeren Werten im Vergleich zum QLQ-F17.
- Für Item Nr. 14 (soziale Funktion) des QLQ-F17 konnte eine nicht-uniforme DIF festgestellt werden: Personen mit geringer Ausprägung des latenten Merkmals erzielten tendenziell höhere Werte im QLQ-F17, während bei hoher Merkmalsausprägung geringere Werte im Vergleich zum QLQ-C30 berichtet wurden.

***Tabelle 11**: Art und Interpretation der detektierten DIF-Items*

Item # F17	Item # C30	Skala	DIF-Typ	Interpretation
8	20	CF	Uniforme DIF	QLQ-F17 weniger funktionelle Einschränkung als QLQ-C30
9	21	EF	Uniforme DIF	QLQ-F17 höhere funktionelle Einschränkung als QLQ-C30
10	22	EF	Uniforme DIF	QLQ-F17 höhere funktionelle Einschränkung als QLQ-C30
14	26	SF	Nicht-uniforme DIF	Bei niedriger Ausprägung: QLQ-F17 höhere funktionelle Einschränkung als QLQ-C30; bei hoher Ausprägung: QLQ-F17 identisch zu QLQ-C30
15	27	SF	Uniforme DIF	QLQ-F17 höhere funktionelle Einschränkung als QLQ-C30

CF, kognitive Funktion; EF, emotionale Funktion; SF, soziale Funktion

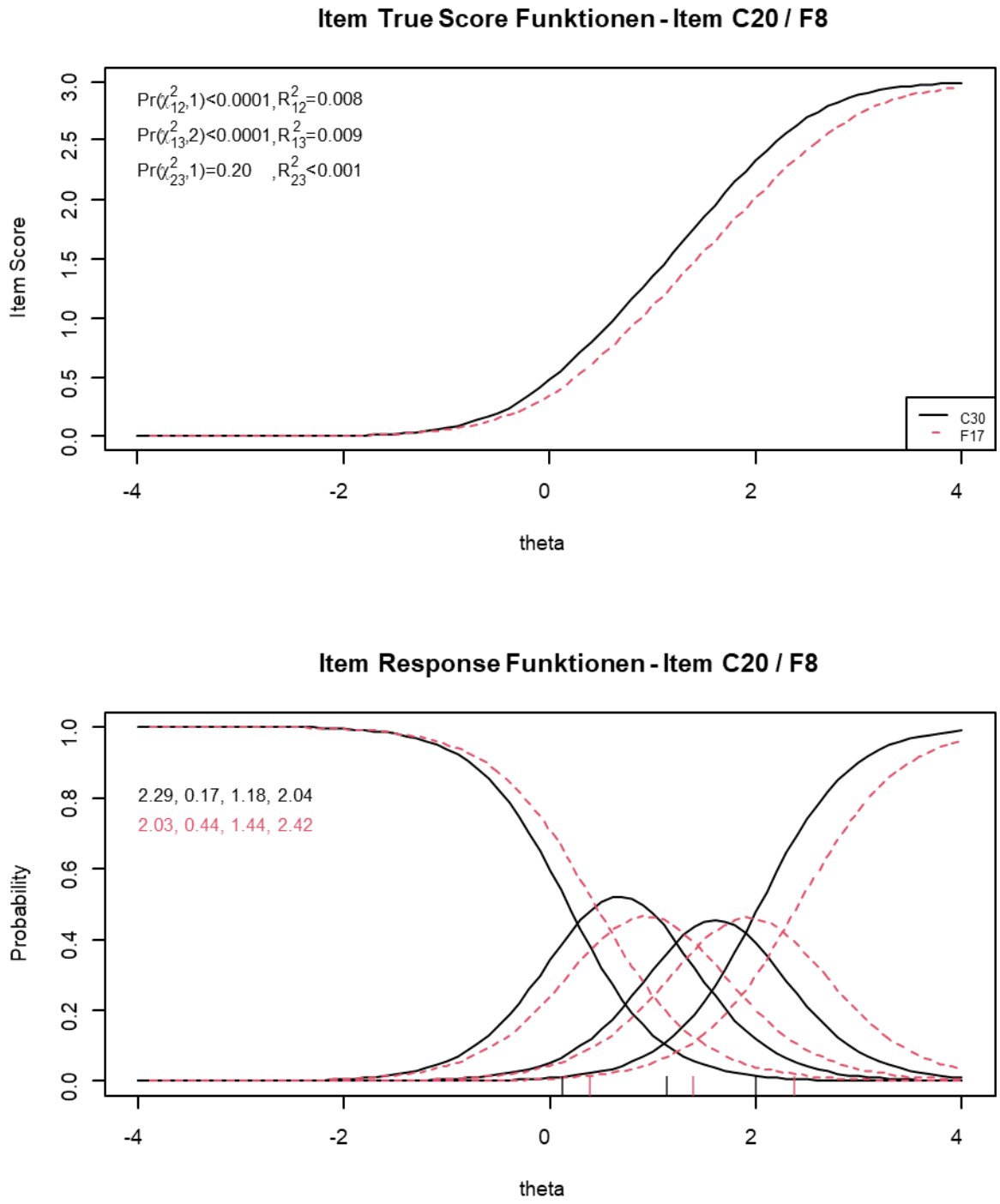


***Abbildung 12**: Diagnostische Diagramme für das DIF-Item C20 / F8. Im oberen Diagramm („Item True Score Funktionen“) sind oben links die p-Werte des Likelihood Ratio χ^2Tests für die Modellvergleiche dargestellt: Modell 1 vs. Modell 2 (uniforme DIF), Modell 1 vs. Modell 3 (allgemeiner Nachweis von DIF) und Modell 2 vs. Modell 3 (nicht-uniforme DIF), jeweils ergänzt um die zugehörigen Pseudo-R^2 Werte. Im unteren Diagramm („Item Response Funktionen“) sind als Kenngrößen die Steigung (Slope) des logistischen Regressionsmodells sowie die Schwellenwerte der jeweiligen Antwortkategorien angegeben (1).*

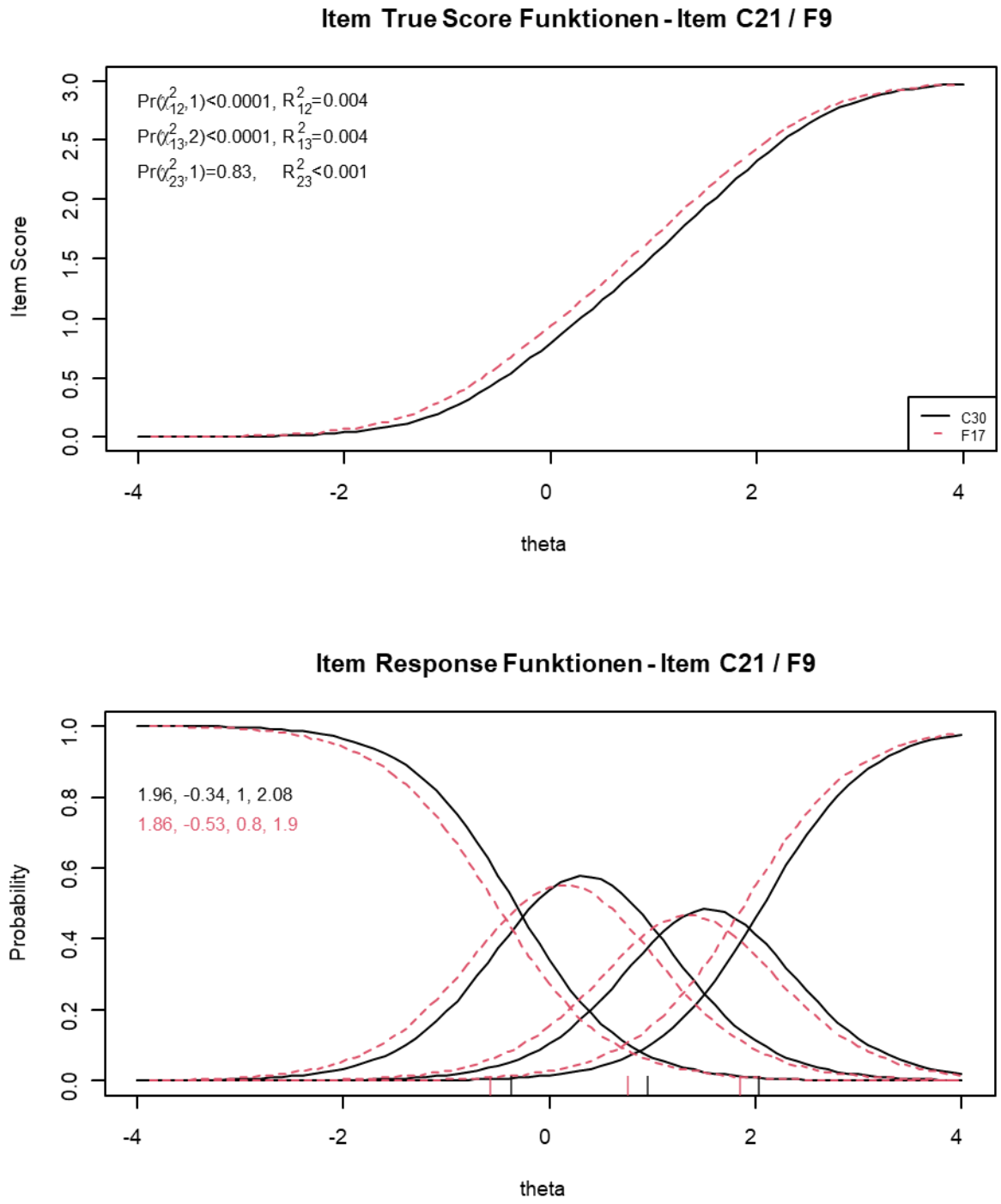


***Abbildung 13**: Diagnostische Diagramme für das DIF-Item C21 / F9. Im oberen Diagramm („Item True Score Funktionen“) sind oben links die p-Werte des Likelihood Ratio χ^2Tests für die Modellvergleiche dargestellt: Modell 1 vs. Modell 2 (uniforme DIF), Modell 1 vs. Modell 3 (allgemeiner Nachweis von DIF) und Modell 2 vs. Modell 3 (nicht-uniforme DIF), jeweils ergänzt um die zugehörigen Pseudo-R^2 Werte. Im unteren Diagramm („Item Response Funktionen“) sind als Kenngrößen die Steigung (Slope) des logistischen Regressionsmodells sowie die Schwellenwerte der jeweiligen Antwortkategorien angegeben (1).*

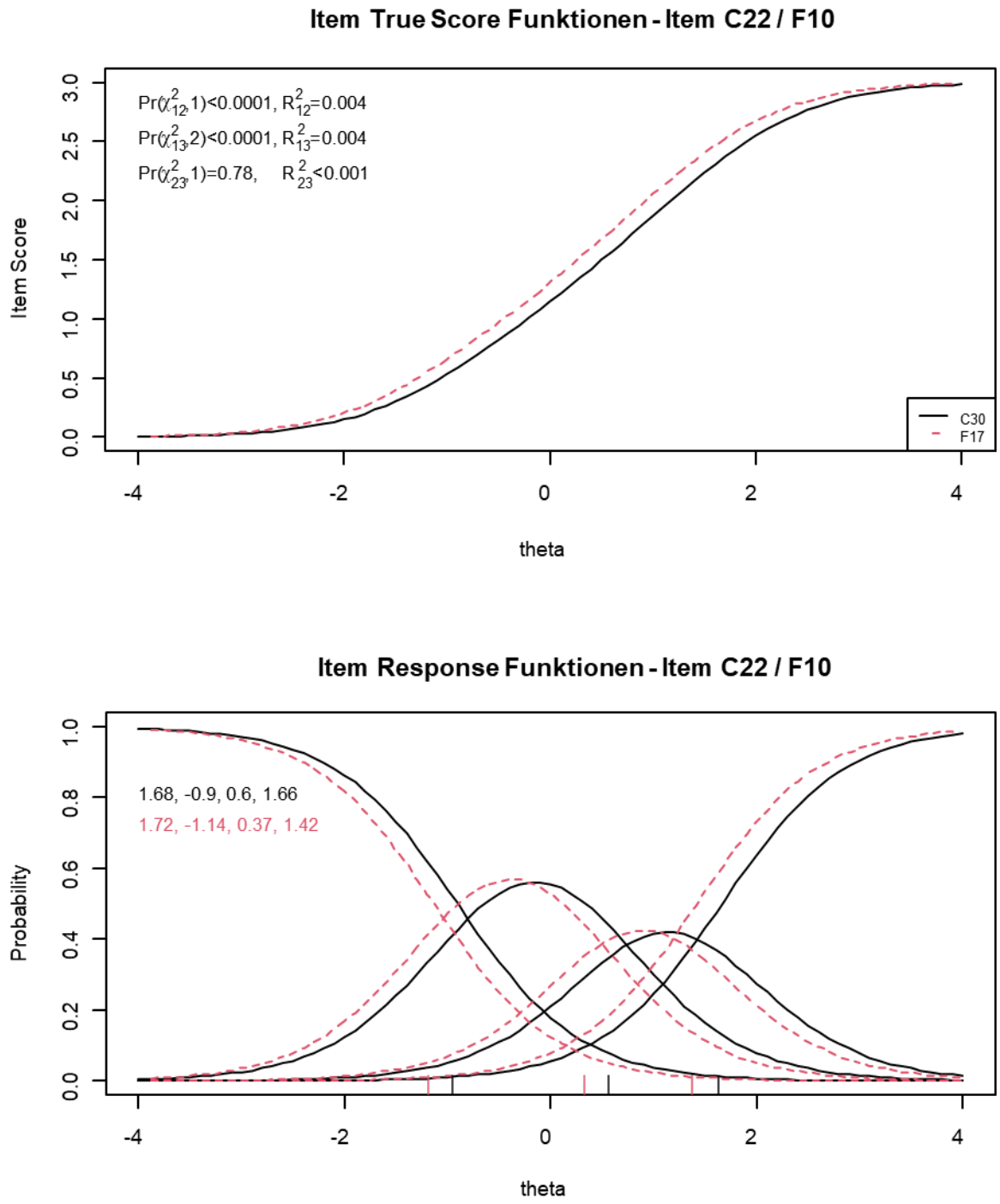


***Abbildung 14**: Diagnostische Diagramme für das DIF-Item C22 / F10. Im oberen Diagramm („Item True Score Funktionen“) sind oben links die p-Werte des Likelihood Ratio χ^2Tests für die Modellvergleiche dargestellt: Modell 1 vs. Modell 2 (uniforme DIF), Modell 1 vs. Modell 3 (allgemeiner Nachweis von DIF) und Modell 2 vs. Modell 3 (nicht-uniforme DIF), jeweils ergänzt um die zugehörigen Pseudo-R^2 Werte. Im unteren Diagramm („Item Response Funktionen“) sind als Kenngrößen die Steigung (Slope) des logistischen Regressionsmodells sowie die Schwellenwerte der jeweiligen Antwortkategorien angegeben (1).*

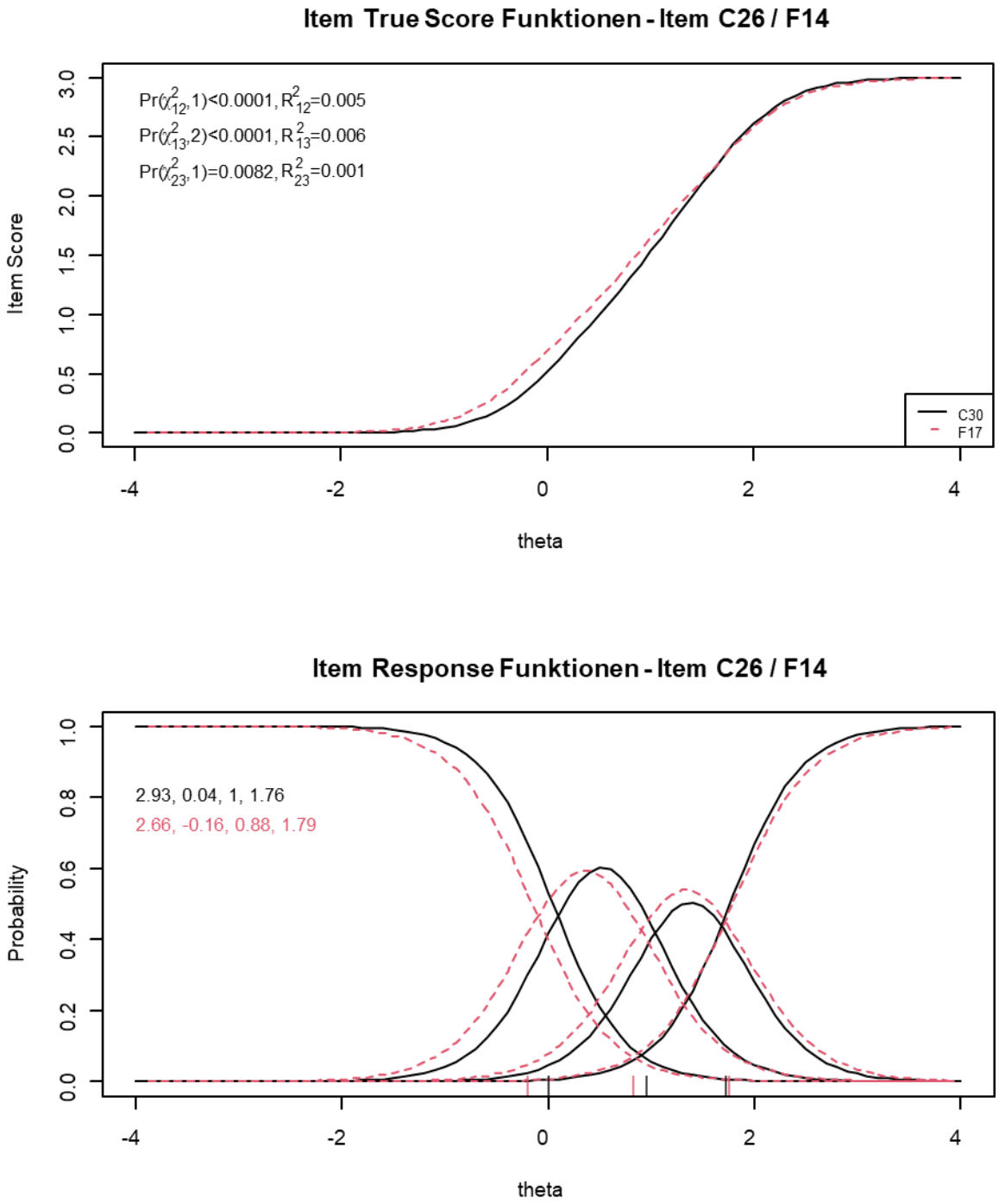


***Abbildung 15**: Diagnostische Diagramme für das DIF-Item C26 / F14. Im oberen Diagramm („Item True Score Funktionen“) sind oben links die p-Werte des Likelihood Ratio χ^2Tests für die Modellvergleiche dargestellt: Modell 1 vs. Modell 2 (uniforme DIF), Modell 1 vs. Modell 3 (allgemeiner Nachweis von DIF) und Modell 2 vs. Modell 3 (nicht-uniforme DIF), jeweils ergänzt um die zugehörigen Pseudo-R^2 Werte. Im unteren Diagramm („Item Response Funktionen“) sind als Kenngrößen die Steigung (Slope) des logistischen Regressionsmodells sowie die Schwellenwerte der jeweiligen Antwortkategorien angegeben (1).*

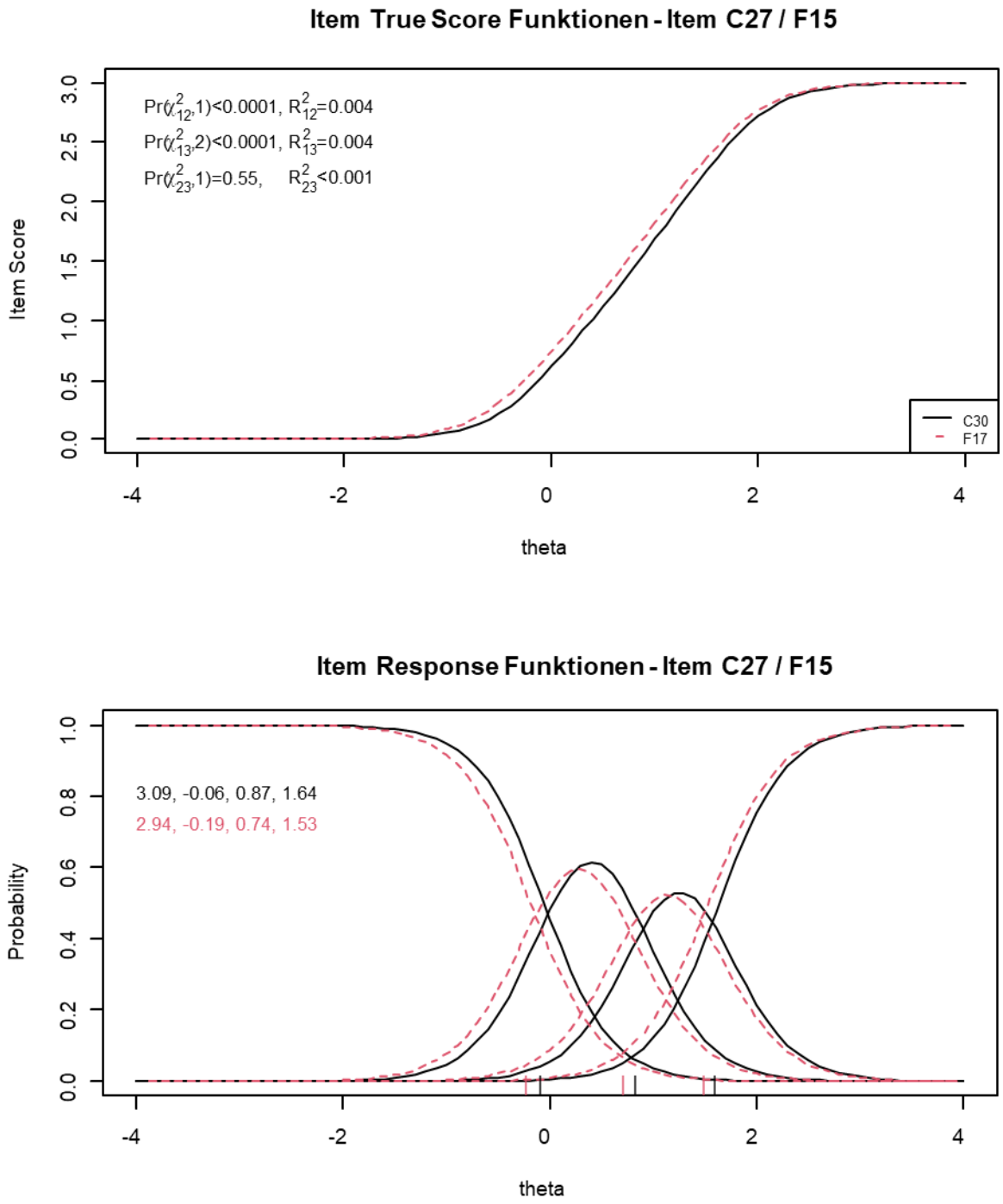


***Abbildung 16**: Diagnostische Diagramme für das DIF-Item C27 / F15. Im oberen Diagramm („Item True Score Funktionen“) sind oben links die p-Werte des Likelihood Ratio χ^2Tests für die Modellvergleiche dargestellt: Modell 1 vs. Modell 2 (uniforme DIF), Modell 1 vs. Modell 3 (allgemeiner Nachweis von DIF) und Modell 2 vs. Modell 3 (nicht-uniforme DIF), jeweils ergänzt um die zugehörigen Pseudo-R^2 Werte. Im unteren Diagramm („Item Response Funktionen“) sind als Kenngrößen die Steigung (Slope) des logistischen Regressionsmodells sowie die Schwellenwerte der jeweiligen Antwortkategorien angegeben (1).*

3.3.2 MULTIPLE LINEARE REGRESSIONSANALYSEN

Zur weiteren Prüfung der Äquivalenz wurden multiple lineare Regressionen für jede Skala durchgeführt, wobei der jeweilige Skalenwert als abhängige Variable und der Fragebogentyp (QLQ-C30 oder QLQ-F17) sowie weitere relevante Kovariablen (Alter, Geschlecht, Land, aktueller Krebsstatus, Q168, aktuelle Behandlung und Aktivitätsniveau) als Prädiktoren in das Modell aufgenommen wurden. Der geschätzte Beta-Koeffizient für den Prädiktor „Fragebogentyp“ gibt dabei die mittlere Differenz der Skalenwerte zwischen den beiden Versionen an. Die mittleren Differenzen sowie die zugehörigen 95 %-Konfidenzintervalle aller Skalen lagen vollständig innerhalb der vorab definierten Äquivalenzgrenzen von]–5; 5[, was den statistischen Nachweis der Äquivalenz der Skalenwerte zwischen den beiden Fragebögen belegt. Diese Ergebnisse stimmen mit den Befunden aus den DIF-Analysen auf Item-Ebene überein und zeigen konsistent die größten, jedoch klinisch nicht relevanten Unterschiede bei den Skalen emotionale Funktion, soziale Funktion und kognitive Funktion. Die vollständigen Ergebnisse der Regressionsanalysen sowie die grafische Darstellung der mittleren Differenzen pro Skala sind in Abbildung 17 zusammengefasst.

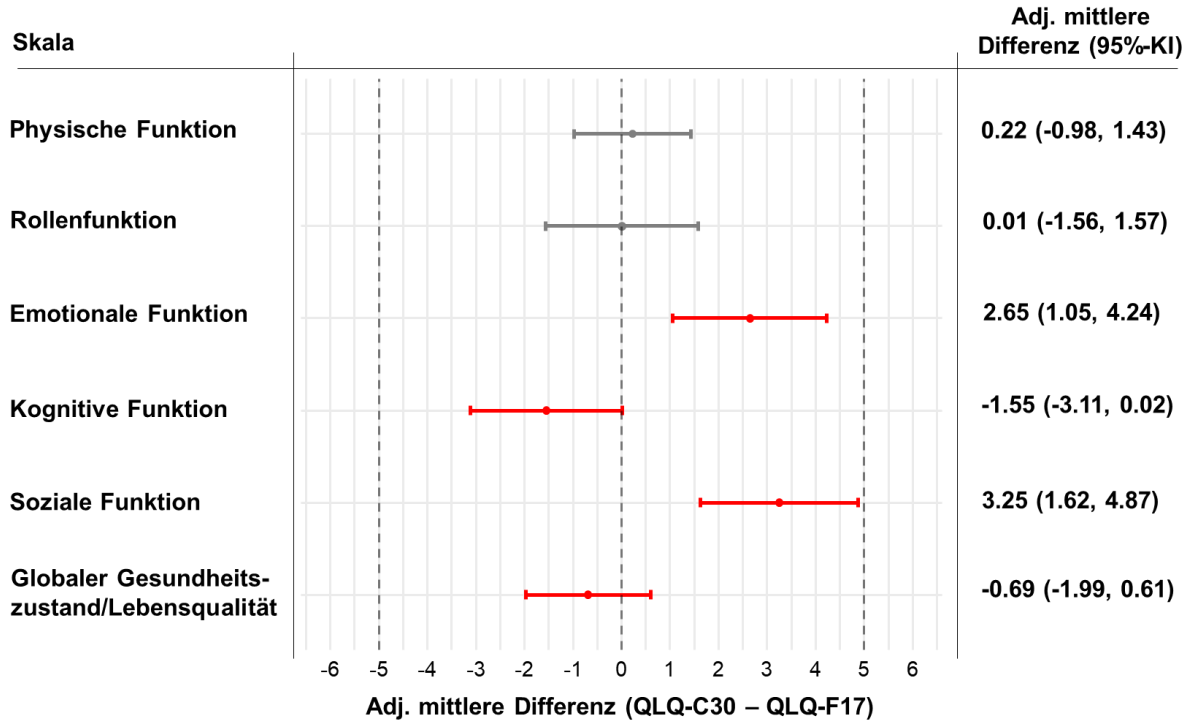


***Abbildung 17**: Forest Plot der multiplen linearen Regression im Zwischen-Gruppenvergleich. Die Grafik zeigt die adjustieren mittleren Differenzen der multiplen linearen Regressionsmodelle, die die Skalen des QLQ-F17 und QLQ-C30 der ersten*

Erhebung vergleichen. Die linearen Modelle wurden adjustiert nach Alter, Geschlecht („Männer“, „Frauen“), Land („Australien“, „Deutschland“, „Finnland“, „Frankreich“, „Italien“, „Polen“, „Rumänien“, „Schweden“, „Spanien“, „Vereinigtes Königreich“, „Vereinigte Staaten von Amerika“), Krebsstatus („Bei mir wurde in den letzten 3 Monaten zum ersten Mal Krebs diagnostiziert.“, „Ich befinde mich derzeit in Krebstherapie.“, „Ich bin in Remission / Ich habe meinen Krebs besiegt.“), Q168 („In welchem Ausmaß fühlten Sie sich durch Nebenwirkungen Ihrer Behandlung belästigt?“), Art der aktiven Behandlung („Chemotherapie“, „Immun- oder zielgerichtete Therapie“, „Radiotherapie“, „Operation innerhalb der letzten 3 Monate“, „andere Therapie“, „Keine aktive Behandlung“) und Aktivitätsgrad („Voll aktiv und fähig, alle Tätigkeiten ohne Einschränkung durchzuführen“, „Aktiv, mit leichter Einschränkung bei körperlich anstrengenden Tätigkeiten“, „Einschränkungen in der Aktivität und bei körperlich anstrengenden Tätigkeiten“, „Selbstversorgung möglich, aber nicht fähig, einer Arbeitstätigkeit nachzugehen“, „Eingeschränkte Selbstversorgung mit mehr als 50% des Tages im Bett oder im Stuhl“); Insgesamt wurden n = 2.683 Patienten in die Modelle aufgenommen; 9 Patienten wurden aufgrund fehlender Angaben zum Geschlecht ausgeschlossen. Die gepunkteten grauen vertikalen Linien zeigen die Äquivalenzbereich von]−5, 5[an. Graue Skalen bestehen aus Items, die in beiden Fragebögen an derselben Position stehen. Die Positionen der Items der roten Skalen unterscheiden sich zwischen QLQ-C30 und QLQ-F17 (1).

3.4 ÄQUIVALENZ – INNERHALB-GRUPPENVERGLEICHE

3.4.1 LINEARE GEMISCHTE MODELLE

Zur Überprüfung der Äquivalenz der gemeinsamen Skalen der beiden Fragebögen unter Verwendung des direkten Vergleichs innerhalb der Patienten wurden lineare gemischte Modelle eingesetzt, welche Unterschiede zwischen den Fragebogenversionen unter Kontrolle individueller Effekte sowie potenzieller Reihenfolgeeffekte untersuchen. Dabei wurde für jede Skala ein Modell geschätzt, das sowohl den Haupteffekt des Fragebogentyps als auch die Interaktion von Fragebogenversion und Sequenz berücksichtigt.

Die Innerhalb-Gruppenvergleiche zur Analyse der Äquivalenz zeigten für alle Skalen mittlere Differenzen nahe 0 mit engen Konfidenzintervallen. Die oberen und unteren Grenzen der 95%-KIs lagen alle innerhalb der Äquivalenzgrenzen von ±5 Punkten, was die Äquivalenz für alle Skalen bestätigte (siehe Abbildung 18).

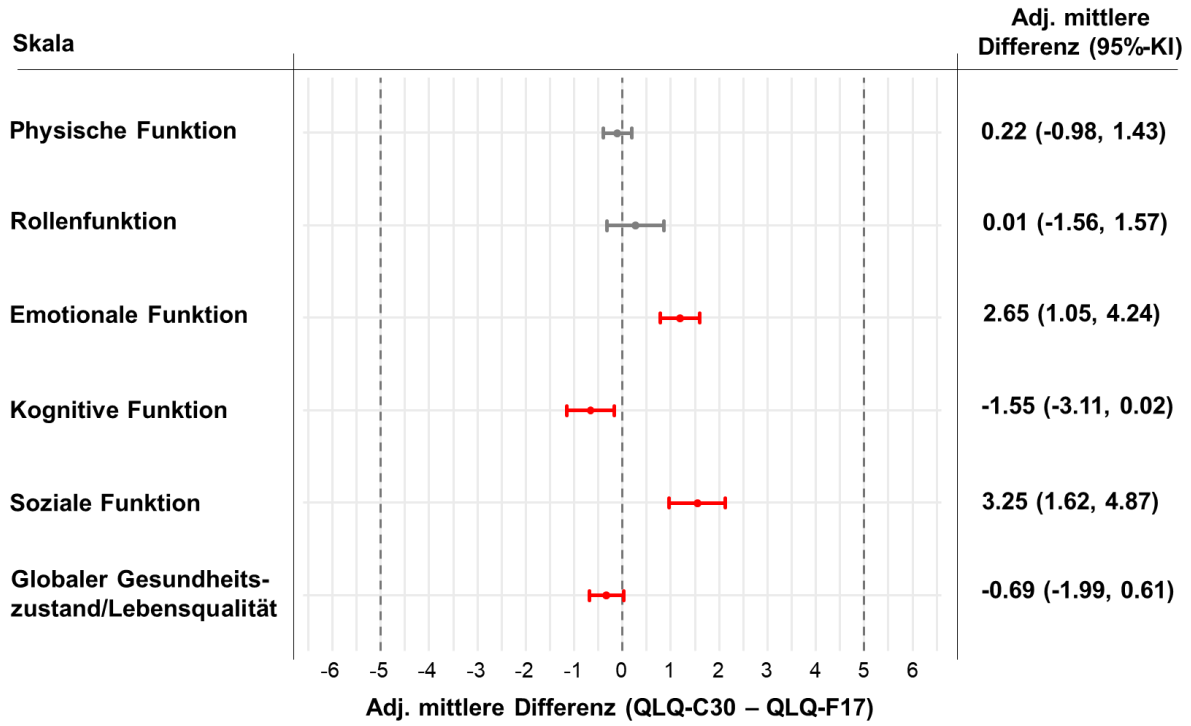


***Abbildung 18**: Forest Plot der linear gemischten Modelle im Innerhalb-Gruppenvergleich. Die Grafik zeigt die mittleren Differenzen der linear gemischten Modelle, die die Skalen des QLQ-F17 und QLQ-C30 innerhalb der Patienten der ersten und zweiten Erhebung im Cross-over Design vergleichen. Die gepunkteten grauen vertikalen Linien zeigen die Äquivalenzbereich von]−5, 5[an. Graue Skalen bestehen aus Items, die in beiden Fragebögen an derselben Position stehen. Die Positionen der Items der roten Skalen unterscheiden sich zwischen QLQ-C30 und QLQ-F17 (1).*

Der Interaktionseffekt zwischen Fragebogenversion und Sequenz zeigte für keine der Skalen einen signifikanten Effekt. Für die Skalen EF, CF und SF wurde jedoch bei der ersten Erhebung ein kleiner Unterschied zwischen dem QLQ-F17 und dem QLQ-C30 beobachtet, während beide Fragebögen bei der zweiten Erhebung nahezu identische Ergebnisse zeigten, was auf leichte Carry-over-Effekte hindeutet (Abbildung 19) (66).

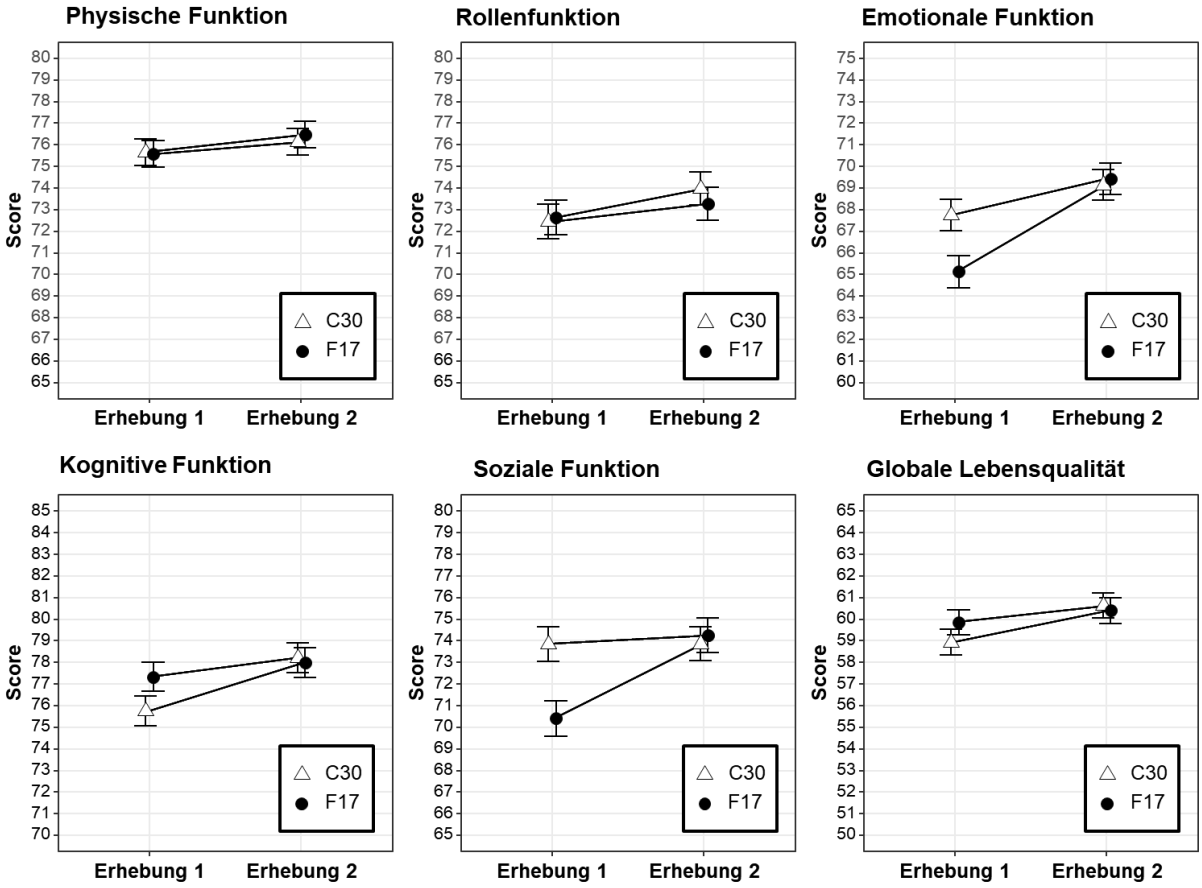


***Abbildung 19**: Mittelwerte und 95%-Konfidenzintervalle der Skalen des QLQ-F17 und des QLQ-C30 in Abhängigkeit vom Zeitpunkt der Erhebung (1).*

3.4.2 UNTERSTÜTZENDE ANALYSEN

Der Prozentsatz der exakten Übereinstimmung für jedes Item (gleiche Antwort in beiden Fragebögen) lag für die ersten 7 Items zwischen 75 % und 94 % und für Items 10-17 zwischen 75 % und 83 %. Der Prozentsatz einer maximalen Abweichung von einem Punkt lag für die ersten 7 Items zwischen 98 % und 99 % und für Items 10-17 zwischen 95 % und 98 %. Die gewichteten Kappa-Koeffizienten waren für alle Items ähnlich und lagen für die ersten 7 Items zwischen 0,62 und 0,75 und für Items 10-17 zwischen 0,63 und 0,75 (siehe Tabelle 12). Auch hier zeigt sich für Frage 8 des QLQ-F17 der kleinste gewichtete Kappa-Wert, was die bereits in den Analysen zuvor entdeckten Effekte unterstreicht.

***Tabelle 12**: Übereinstimmung auf Item-Ebene und gewichtetes Kappa (Test-Retest-Reliabilität auf Item-Ebene) (1)*

Item # F17	Item # C30	Skala	Absolute Übereinstimmung (95%-KI)	≤1 Abweichung* (95%-KI)	Gewichtetes Kappa (95%-KI)
1	1	PF	79% (77%, 80%)	99% (98%, 99%)	0.72 (0.69, 0.74)
2	2	PF	81% (79%, 82%)	99% (98%, 99%)	0.75 (0.72, 0.78)
3	3	PF	88% (86%, 89%)	99% (98%, 99%)	0.66 (0.62, 0.71)
4	4	PF	83% (81%, 84%)	98% (98%, 99%)	0.64 (0.60, 0.68)
5	5	PF	94% (92%, 94%)	99% (98%, 99%)	0.60 (0.53, 0.68)
6	6	RF	75% (74%, 77%)	98% (97%, 99%)	0.59 (0.55, 0.62)
7	7	RF	76% (74%, 77%)	98% (97%, 98%)	0.61 (0.58, 0.65)
8	20	CF	79% (77%, 81%)	97% (97%, 98%)	0.54 (0.49, 0.59)
9	21	EF	76% (74%, 77%)	98% (97%, 98%)	0.61 (0.57, 0.64)
10	22	EF	76% (74%, 77%)	97% (96%, 98%)	0.71 (0.68, 0.74)
11	23	EF	83% (82%, 85%)	99% (99%, 99%)	0.75 (0.72, 0.78)
12	24	EF	82% (80%, 83%)	98% (98%, 99%)	0.72 (0.69, 0.75)
13	25	CF	81% (80%, 83%)	98% (98%, 99%)	0.67 (0.63, 0.71)
14	26	SF	76% (75%, 78%)	97% (96%, 98%)	0.57 (0.53, 0.61)
15	27	SF	78% (76%, 79%)	97% (97%, 98%)	0.64 (0.61, 0.68)
16	19	QL	76% (74%, 78%)	95% (94%, 96%)	0.72 (0.69, 0.74)
17	30	QL	75% (73%, 76%)	95% (95%, 96%)	0.75 (0.72, 0.78)

** Abweichend um höchstens eine Antwortkategorie im Sinne der ordinalen Skala; PF, Physische Funktion; RF, Rollenfunktion; EF, emotionale Funktion; CF, kognitive Funktion; SF, soziale Funktion; QL, globaler Gesundheitszustand/ Lebensqualität.*

Die ICCs waren für alle Skalen hoch (>0,8). Die ICCs der Skalen von Block 2 lagen zwischen den ICCs von Block 1 (siehe Abbildung 20).

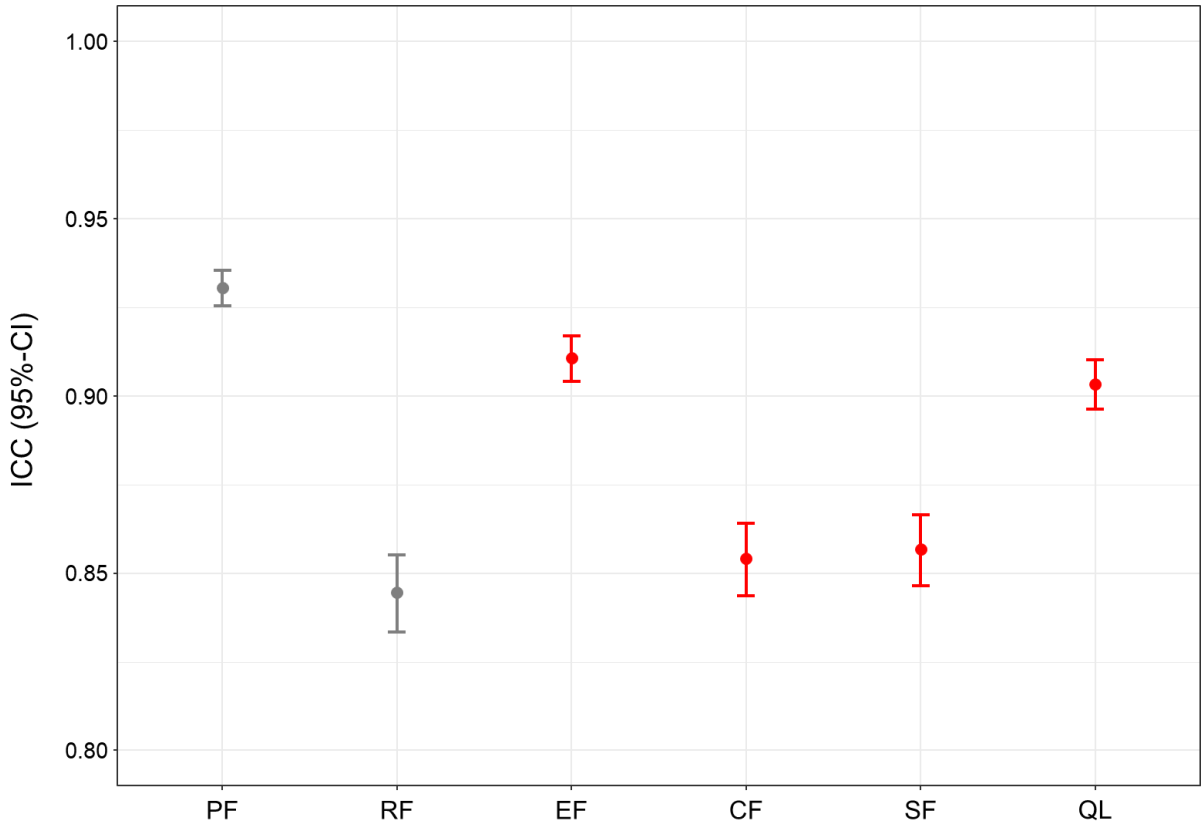


***Abbildung 20**: Intraklassen-Korrelationkoeffizienten (ICC) inkl. 95%-Konfidenz-intervall zwischen dem QLQ-C30 und dem QLQ-F17 für jede Skala. PF, Physische Funktion; RF, Rollenfunktion; EF, emotionale Funktion; CF, kognitive Funktion; SF, soziale Funktion; QL, globaler Gesundheitszustand/ Lebensqualität (1).*

3.5 SUBGRUPPENANALYSEN

3.5.1 SYMPTOMLAST

Die DIF-Analysen in den symptombezogenen Subgruppen zeigten ein differenziertes Bild in Abhängigkeit von der individuellen Symptomlast der Patienten.

In der Subgruppe mit niedriger Symptomlast (Summenwert der acht Symptomskalen < 200) wurden die Items 9 („Fühlten Sie sich angespannt?“), 10 („Haben Sie sich Sorgen gemacht?“), 14 („Hat Ihr körperlicher Zustand oder Ihre medizinische Behandlung Ihr Familienleben beeinträchtigt?“) und 15 („Hat Ihr körperlicher Zustand oder Ihre medizinische Behandlung Ihr Zusammensein oder Ihre gemeinsamen Unternehmungen mit anderen Menschen beeinträchtigt?“) als DIF-Items identifiziert. Auffällig war, dass Item 8 („Hatten Sie Schwierigkeiten, sich auf etwas zu konzentrieren, z. B. auf Zeitunglesen oder Fernsehen?“), welches in der

Gesamtkohorte einen der stärksten Effekte zeigte, in dieser Subgruppe keine DIF aufwies.

In der Subgruppe mit mittlerer Symptomlast (Summenwert 200–400) zeigte sich hingegen ein DIF-Muster, das weitgehend dem der Gesamtkohorte entsprach. Zusätzlich wurde hier Item 13 („Waren Sie durch Ihre körperlichen Beschwerden beim Essen oder Trinken eingeschränkt?“) als weiteres DIF-Item identifiziert. Dies deutet darauf hin, dass mit zunehmender Symptomlast weitere Items sensitiv gegenüber Reihenfolgeeffekten werden könnten.

Bei Patienten mit hoher Symptomlast (Summenwert > 400) zeigte sich ausschließlich bei Item 8 ein signifikanter DIF-Effekt, mit einer Effektstärke von 0,042, die im Vergleich zu den übrigen Subgruppen als auffällig hoch einzustufen ist. Aufgrund der geringen Fallzahl in dieser Subgruppe (n=293) ist es möglich, dass weitere bestehende DIF-Effekte nicht detektiert werden konnten (52). Dennoch deutet sich ein Zusammenhang zwischen hoher Symptomlast und einem verstärkten Reihenfolgeeffekt bei Item 8 an, das direkt im Anschluss an die Symptomskalen platziert ist.

In keiner der drei Subgruppen konnte nicht-uniforme DIF identifiziert werden. Die Ergebnisse der signifikanten DIF-Items in den jeweiligen Subgruppen sind in Tabelle 13 dargestellt.

***Tabelle 13**: Signifikante DIF-Items in Abhängigkeit der Symptomlast*

Item # F17	Item # C30	Skala	%Beta	Effekt* uniform	p-Wert uniform	Effekt* nicht-uniform	p-Wert nicht-uniform
Niedrige Symptomlast (n=1.582) (QLQ-C30: n=751, QLQ-F17: n=831)							
9	21	PF	0,006	0,013	0,000	0,002	0,091
10	22	CF	0,004	0,012	0,000	0,002	0,048
14	26	SF	0,007	0,011	0,000	0,000	0,830
15	27	SF	0,006	0,007	0,000	0,001	0,116

Mittlere Symptomlast (n=768) (QLQ-C30: n=400, QLQ-F17: n=368)							
8	20	CF	0,037	0,012	0,002	0,001	0,299
9	21	EF	0,032	0,015	0,001	0,001	0,529
10	22	EF	0,038	0,025	0,000	0,000	0,675
13	25	CF	0,018	0,012	0,003	0,000	0,853
14	26	SF	0,004	0,017	0,000	0,000	0,781
15	27	SF	0,001	0,007	0,002	0,000	0,535
Hohe Symptomlast (n=293) (QLQ-C30: n=169, QLQ-F17: n=124)							
8	20	CF	0,085	0,042	0,000	0,001	0,563

**Pseudo-R² Werte nach Nagelkerke; %Beta, relative Veränderung des Beta-Wertes von Modell 1 zu Modell 2; PF, Physische Funktion; EF, emotionale Funktion; CF, kognitive Funktion; SF, soziale Funktion; Die Tabelle zeigt nur die identifizierten DIF-Items. Die Ergebnisse beziehen sich auf die erste Erhebung und vergleichen Patienten unterteilt in Subgruppen nach Symptomlast, die entweder den EORTC QLQ-C30 oder den EORTC QLQ-F17 ausgefüllt haben. Niedrige Symptomlast (grüne Zeilen) ist definiert durch einen Summenwert aller Symptomskalen <200; Mittlere Symptomlast (gelbe Zeilen) ist definiert durch einen Summerwert aller Symptomskalen zwischen 200 und 400. Hohe Symptomlast (rote Zeilen) ist definiert durch einen Summerwert aller Symptomskalen >400. Ein p-Wert < 0,01 weist auf eine uniforme oder nicht-uniforme DIF einer Frage hin. Eine Effektgröße < 0,01 kann als trivial und nicht aussagekräftig angesehen werden.*

3.5.2 Geschlecht

Die geschlechtsspezifischen DIF-Analysen zeigten bei den weiblichen Patientinnen ein identisches Muster zu jenem der Gesamtkohorte. Alle Items, bei denen in der Hauptanalyse DIF-Effekte identifiziert wurden, zeigten sich auch in der Subgruppe der Frauen mit vergleichbarer Richtung und Größenordnung. Bei den männlichen Patienten hingegen fiel auf, dass für die beiden Items zur sozialen Funktion kein signifikanter DIF-Effekt detektiert wurde. Dafür zeigte sich bei den Männern bei Item 3 („Benötigten Sie Hilfe beim Essen, Anziehen, Waschen oder Benutzen der Toilette?“) ein leichter DIF-Effekt, der jedoch aufgrund seiner geringen Effektstärke sowie fehlender inhaltlicher Plausibilität als statistisches Artefakt interpretiert wird.

Insgesamt lagen alle geschätzten DIF-Effekte in beiden Subgruppen im sehr kleinen Bereich. Eine Ausnahme bildete lediglich Item 8 („Hatten Sie Schwierigkeiten, sich auf etwas zu konzentrieren, z. B. auf Zeitunglesen oder Fernsehen?“), das bei männlichen Patienten mit einer Effektgröße von 0,012 einen leicht erhöhten Wert im Vergleich zur Gesamtkohorte zeigte. Nicht-uniforme DIF-Effekte konnten in keiner der beiden Geschlechts-Subgruppen identifiziert werden. Alle identifizierten DIF-Items in den beiden Subgruppen sind in Tabelle 14 dargestellt.

***Tabelle 14**: Signifikante DIF-Items der Subgruppen Männer und Frauen*

Item # F17	Item # C30	Skala	%Beta	Effekt* uniform	p-Wert uniform	Effekt* nicht-uniform	p-Wert nicht-uniform
Männer (n=1.311) (QLQ-C30: n=650, QLQ-F17: n=661)							
3	3	PF	0,009	0,005	**0,002**	<0,001	0,335
8	20	CF	0,023	0,012	**<0,001**	<0,001	0,522
9	21	EF	0,005	0,004	**0,002**	<0,001	0,513
10	22	EF	0,004	0,005	**0,001**	<0,001	0,806
Frauen (n=1.323) (QLQ-C30: n=664, QLQ-F17: n=659)							
8	20	CF	0,009	0,008	**<0,001**	0,001	0,282
9	21	EF	0,008	0,005	**0,001**	<0,001	0,762
10	22	EF	0,008	0,006	**<0,001**	<0,001	0,436
14	26	SF	0,016	0,007	**<0,001**	0,001	0,039
15	27	SF	0,009	0,003	**0,001**	<0,001	0,570

**Pseudo-R^2 Werte nach Nagelkerke; %Beta, relative Veränderung des Beta-Wertes von Modell 1 zu Modell 2; PF, Physische Funktion; EF, emotionale Funktion; CF, kognitive Funktion; SF, soziale Funktion; Die Tabelle zeigt nur die identifizierten DIF-Items. Die Ergebnisse beziehen sich auf die erste Erhebung und vergleichen sowohl Männer (blau) als auch Frauen (gelb), die entweder den EORTC QLQ-C30 oder den EORTC QLQ-F17 ausgefüllt haben. Ein p-Wert < 0,01 weist auf eine uniforme oder nicht-uniforme DIF einer Frage hin. Eine Effektgröße < 0,01 kann als trivial und nicht aussagekräftig angesehen werden.*

3.6 PSYCHOMETRISCHE EIGENSCHAFTEN DES QLQ-F17

Zusätzlich zur Äquivalenz des QLQ-F17 zu den entsprechenden Skalen des QLQ-C30 mussten auch die grundlegenden psychometrischen Eigenschaften des QLQ-F17 im Vergleich zum QLQ-C30 analysiert werden. Erst durch die Bestätigung der Faktorenstruktur, der internen Konsistenz und der Validität des QLQ-F17 gilt dieser als vollständig validiert.

Konfirmatorische Faktorenanalyse (KFA)

Im ersten Schritt wurde die Skalenstruktur des QLQ-F17 mit einer konfirmatorischen Faktorenanalyse analysiert. Dabei wurde ein Sechs-Faktoren-Modell spezifiziert, das analog zum QLQ-C30 die Dimensionen körperliche Funktion, Rollenfunktion, emotionale Funktion, kognitive Funktion, soziale Funktion sowie globaler Gesundheitszustand/Lebensqualität umfasst. Die Items wurden entsprechend ihrer Zuordnung im QLQ-C30 auf die jeweiligen latenten Faktoren geladen. Alle Faktorladungen im Modell des QLQ-F17 lagen über dem konventionellen Schwellenwert von 0,40 (siehe Tabelle 15), was auf eine ausreichende konvergente Validität der Skalenstruktur hinweist. Die Fit-Indizes zeigten eine insgesamt akzeptable Modellgüte (siehe Tabelle 16). Die Interfaktorkorrelationen bewegten sich zwischen 0,44 und 0,84 (sieheTabelle 17), was die theoretisch erwartete Zusammengehörigkeit der Skalen unterstützt, ohne auf problematische Multikollinearität hinzudeuten.

***Tabelle 15**: Standardisierte Faktorladungen der konfirmatorischen Faktorenanalyse des EORTC QLQ-F17 (N = 1.323, erste Erhebung) und des QLQ-C30 (N = 1.320, erste Erhebung) (1)*

Skala	Frage	Standardisierte Faktorladungen QLQ-F17	Standardisierte Faktorladungen QLQ-C30
Körperliche Funktion	F1	0,781	0,736
	F2	0,796	0,798
	F3	0,810	0,826
	F4	0,712	0,738
	F5	0,535	0,541

Rollenfunktion	F6	0,880	0,876
	F7	0,866	0,866
Emotionale Funktion	F9	0,829	0,825
	F10	0,829	0,814
	F11	0,784	0,785
	F12	0,833	0,827
Kognitive Funktion	F8	,802	0,816
	F13	0,725	0,689
Soziale Funktion	F14	0,845	0,867
	F15	0,890	0,885
Globaler Gesundheitszustand/ Lebensqualität	F16	0,865	0,875
	F17	0,885	0,907

***Tabelle 16**: Modell-Fit-Indizes der konfirmatorischen Faktorenanalyse des QLQ-F17 (N = 1.323, erste Erhebung) und des QLQ-C30 (N = 1.320, erste Erhebung)*

Fit-Index	Fragebogen	Wert	Grenzwert für gutes Modell	Interpretation
CFI	QLQ-F17	0,941	≥ 0,85	Gutes Modell
	QLQ-C30	0,951		Gutes Modell
TLI	QLQ-F17	0,922	≥ 0,85	Gutes Modell
	QLQ-C30	0,937		Gutes Modell
RMSEA	QLQ-F17	0,079	< 0,08	Akzeptabler Fit
	QLQ-C30	0,072		Akzeptabler Fit
SRMR	QLQ-F17	0,050	< 0,08	Gutes Modell
	QLQ-C30	0,043		Gutes Modell

CFI, Comparative Fit Index (Vergleichsindex der Modellanpassung); TLI, Tucker-Lewis Index; RMSEA, Root Mean Square Error of Approximation (Wurzel des mittleren quadratischen Näherungsfehlers); SRMR, Standardized Root Mean Square Residual

***Tabelle 17**: Interfaktorkorrelationen der konfirmatorischen Faktorenanalyse des EORTC QLQ-F17 (N = 1.323, erste Erhebung) und des QLQ-C30 (N = 1.320, erste Erhebung)*

Skala	PF	RF	EF	CF	SF	QL
PF	1,00					
RF	0,84 (0,86)	1,00				
EF	0,51 (0,55)	0,61 (0,63)	1,00			
CF	0,62 (0,70)	0,68 (0,70)	0,77 (0,85)	1,00		
SF	0,74 (0,72)	0,82 (0,83)	0,71 (0,74)	0,73 (0,79)	1,00	
QL	0,58 (0,58)	0,60 (0,58)	0,54 (0,56)	0,44 (0,52)	0,59 (0,58)	1,00

Die Daten zeigen die Interfaktorkorrelationen des QLQ-F17 (QLQ-C30); PF, Körperliche Funktion; RF, Rollenfunktion; EF, emotionale Funktion; CF, kognitive Funktion; SF, soziale Funktion; QL, globaler Gesundheitszustand/Lebensqualität

Interne Konsistenz

Zur Bewertung der internen Konsistenz wurden für alle Skalen des QLQ-F17 sowie für die analogen Skalen des QLQ-C30 Cronbach's Alpha-Koeffizienten berechnet. Die Ergebnisse zeigten durchgängig sehr gute Werte im Bereich von 0,73 bis 0,89, wobei alle Skalen den etablierten Schwellenwert von $\alpha > 0{,}70$ überschritten. Die Werte lagen dabei in einem ähnlichen Bereich wie jene des QLQ-C30, was auf eine vergleichbare interne Konsistenz beider Instrumente hinweist (siehe Tabelle 18).

Konvergente und diskriminante Validität

Mit Hilfe von Item-Skalen Korrelationen wurden die konvergente Validität und die diskriminante Validität untersucht. Dabei zeigten die Analysen substanzielle Korrelationen aller Items des QLQ-F17 mit ihren zugehörigen Skalen ($r > 0{,}4$), womit die konvergente Validität aller Items gezeigt werden konnte. Skalierungsfehler, d. h. höhere Korrelationen eines Items mit einer nicht zugeordneten Skala als mit der eigenen, traten lediglich bei zwei Items der Skala „körperliche Funktion" und einem Item der Skala „kognitive Funktion" auf. Diese Korrelationen waren nur sehr geringfügig höher (Δ=0,01 bis 0,03) und statistisch nicht signifikant, sodass auch die diskriminante Validität weitgehend bestätigt werden konnte (siehe Tabelle 19).

***Tabelle 18**: Interne Konsistenz, konvergente Validität, diskriminante Validität und Skalierungsfehler des QLQ-F17 (N = 1.323, erste Erhebung) und des QLQ-C30 (N = 1.320, erste Erhebung) (1)*

Skala	Frage-bogen	Cronbach's Alpha (95%-KI)	Konvergente Validität	Diskriminante Validität	Skalierungs-fehler
PF	F17	0,84 (0,83, 0,86)	0,47-0,76	0,38-0,70	2 (8%)
	C30	0,84 (0,83, 0,86)	0,48-0,78	0,17-0,63	0 (0%)
RF	F17	0,86 (0,85, 0,88)	0,76-0,76	0,41-0,68	0 (0%)
	C30	0,86 (0,85, 0,88)	0,76-0,76	0,46-0,72	0 (0%)
EF	F17	0,89 (0,88, 0,90)	0,73-0,78	0,32-0,59	0 (0%)
	C30	0,89 (0,88, 0,90)	0,73-0,76	0,39-0,62	0 (0%)
CF	F17	0,73 (0,70, 0,76)	0,58-0,58	0,32-0,57	1 (10%)
	C30	0,72 (0,69, 0,75)	0,56-0,56	0,32-0,65	2 (20%)
SF	F17	0,86 (0,84, 0,87)	0,75-0,75	0,44-0,66	0 (0%)
	C30	0,87 (0,85, 0,88)	0,77-0,77	0,46-0,69	0 (0%)
QL	F17	0,87 (0,85, 0,88)	0,77-0,77	0,19-0,51	0 (0%)
	C30	0,89 (0,87, 0,90)	0,79-0,79	0,37-0,50	0 (0%)

PF, Körperliche Funktion; RF, Rollenfunktion; EF, emotionale Funktion; CF, kognitive Funktion; SF, soziale Funktion; QL, globaler Gesundheitszustand/Lebensqualität; Konvergente Validität, Korrigierte Item-Skala-Korrelation (Korrelationen von Items mit ihrer eigenen Skala, wobei das jeweilige Item bei der Berechnung der Gesamtskala ausgeschlossen wurde); Diskriminante Validität, Item-Skala-Korrelation mit Fremdskalen; Skalierungsfehler, Anzahl der Fälle, bei denen die Frage mit einer anderen Skala höher korreliert als mit der eigenen Skala

***Tabelle 19**: Korrigierte Item-Skala-Korrelationen des QLQ-F17 (1)*

Skala	Frage	PF	RF	EF	CF	SF	QL
PF	F1	**0,69**	0,62	0,39	0,41	0,53	0,47
	F2	**0,71**	0,58	0,32	0,35	0,50	0,47
	F3	**0,76**	0,61	0,35	0,40	0,51	0,37
	F4	**0,64**	0,59	0,44	0,45	0,58	0,41
	F5	**0,47**	0,50	0,39	0,49	0,44	0,19

RF	F6	0,70	**0,76**	0,49	0,51	0,66	0,51
	F7	0,68	**0,76**	0,51	0,51	0,65	0,47
EF	F9	0,44	0,50	**0,77**	0,56	0,53	0,39
	F10	0,38	0,47	**0,78**	0,50	0,54	0,40
	F11	0,41	0,41	**0,73**	0,54	0,52	0,37
	F12	0,41	0,47	**0,76**	0,57	0,56	0,47
CF	F8	0,48	0,52	0,59	**0,58**	0,54	0,32
	F13	0,45	0,45	0,53	**0,58**	0,50	0,31
SF	F14	0,59	0,63	0,58	0,55	**0,75**	0,46
	F15	0,63	0,68	0,58	0,55	**0,75**	0,49
QL	F16	0,49	0,50	0,40	0,32	0,47	**0,77**
	F17	0,46	0,48	0,48	0,34	0,48	**0,77**

PF, Physische Funktion; RF, Rollenfunktion; EF, emotionale Funktion; CF, kognitive Funktion; SF, soziale Funktion; QL, globaler Gesundheitszustand/Lebensqualität; Die Werte stellen Pearson Korrelations Koeffizienten dar. Fett gedruckte Koeffizienten entsprechen den Korrelationen mit der eigenen Skala; Rot gedruckte Koeffizienten zeigen Korrelationen, die mit einer Fremdskala höher sind als mit der zugehörigen eigenen Skala.

3.7 Repräsentativität der Analysepopulation

Zur Bewertung der Repräsentativität der Analysepopulation hinsichtlich funktionaler und symptomatischer Lebensqualität im Vergleich zur „allgemeinen“ Population von Krebspatienten wurde ein Vergleich der mittleren Skalenwerte des EORTC QLQ-C30 mit den Referenzwerten des EORTC-Referenzhandbuchs durchgeführt (47). Das Referenzhandbuch basiert auf Daten von über 23.000 Krebspatienten aus zahlreichen Ländern, die verschiedene Altersgruppen, Tumorentitäten sowie Krankheitsstadien repräsentieren. Da die funktionale Lebensqualität stark altersabhängig ist, wurde die Analysepopulation auch innerhalb der einzelnen Altersgruppen gegen die Referenzpopulation verglichen. Die Mittelwerte und Standardabweichungen der Funktionsskalen in der vorliegenden Stichprobe zeigten eine sehr gute Übereinstimmung mit den Referenzwerten (siehe Tabelle 20).

Tabelle 20: *EORTC QLQ-C30 Mittelwerte der Funktionsskalen im Vergleich zu Referenzwerten nach Altersgruppen und insgesamt (1)*

Alters-gruppe	Population	PF	RF	EF	CF	SF	QL
<50	Analysepopulation (n=739)*	68,6 ±25,0	63,8 ±29,2	53,1 ±27,4	64,0 ±29,0	61,3 ±30,7	57,5 ±22,2
	Referenz (n=5,207)**	80,2 ±20,8	68,6 ±31,7	69,2 ±24,4	82,9 ±21,6	72,1 ±29,5	61,4 ±23,4
50-59	Analysepopulation (n=484)*	78,3 ±21,6	72,0 ±28,7	64,8 ±26,2	75,5 ±25,4	72,6 ±29,2	58,4 ±21,8
	Referenz (n=5,707)**	78,0 ±22,5	69,4 ±32,7	69,0 ±24,2	83,2 ±21,9	73,5 ±29,4	61,2 ±24,1
60-69	Analysepopulation (n=730)*	78,9 ±20,1	77,2 ±26,5	74,2 ±22,9	83,4 ±20,4	79,3 ±26,0	60,5 ±21,1
	Referenz (n=6,709)**	76,3 ±23,5	72,6 ±32,7	71,8 ±24,3	83,1 ±21,6	76,4 ±28,8	61,8 ±24,4
≥70	Analysepopulation (n=690)*	78,8 ±20,3	79,9 ±25,2	81,3 ±18,6	85,1 ±17,6	82,5 ±23,8	62,5 ±20,6
	Referenz (n=5,357)**	72,1 ±25,4	70,7 ±34,1	76,1 ±23,2	81,0 ±22,4	78,2 ±28,2	60,6 ±25,1
Gesamt	Analysepopulation (n=2.643)*	75,9 ±22,3	73,2 ±28,1	68,5 ±26,3	77,0 ±25,0	73,9 ±28,7	59,8 ±21,5
	Referenz (n=23.553)**	76,7 ±23,2	70,5 ±32,8	71,4 ±24,2	82,6 ±21,9	75,0 ±29,1	61,3 ±24,2

*Daten zeigen Mittelwert±SD; PF, Physische Funktion; RF, Rollenfunktion; EF, emotionale Funktion; CF, kognitive Funktion; SF, soziale Funktion; QL, globaler Gesundheitszustand/ Lebensqualität (QOL); **Referenzpopulation hauptsächlich basierend auf Daten aus klinischen Krebsstudien und epidemiologischen Untersuchungen mit insgesamt 23.553 Fällen (47).*

Für den Vergleich der Symptomskalen wurde die Studienkohorte zusätzlich in Subgruppen zum aktuellen Krebsstatus unterteilt. Hier zeigte sich, dass sie Subgruppe der Patienten in Remission am ehesten mit der Referenzpopulation übereinstimmt.

***Tabelle 21**: EORTC QLQ-C30 Mittelwerte der Symptomskalen im Vergleich zu Referenzwerten nach Krebsstatus und insgesamt (1)*

Population	FA	NV	PA	DY
Subgruppen nach Krebsstatus				
Ich wurde in den letzten 3 Monaten neu mit Krebs diagnostiziert (n=190)	51,5 ±30,3	34,4 ±35,0	46,8 ±34,0	40,0 ±34,5
Ich befinde mich derzeit in Krebsbehandlung (n=488)	48,8 ±26,9	20,6 ±27,1	40,2 ±29,2	34,6 ±29,9
Ich bin in Remission von Krebs / Ich bin ein Krebsüberlebender (n=1,965)	34,7 ±25,0	7,7 ±16,8	30,2 ±27,8	21,4 ±27,6
Analysepopulation (n=2,643)	38,5 ±26,6	12,0 ±22,3	33,2 ±29,0	25,2 ±29,3
Referenz (n=23,553)*	34,6 ±27,8	9,1 ±19,0	27,0 ±29,9	21,0 ±28,4
Population	**SL**	**AP**	**CO**	**DI**
Subgruppen nach Krebsstatus				
Ich wurde in den letzten 3 Monaten neu mit Krebs diagnostiziert (n=190)	50,7 ±35,6	41,2 ±37,3	37,7 ±36,1	28,8 ±34,8
Ich befinde mich derzeit in Krebsbehandlung (n=488)	44,0 ±32,3	27,4 ±30,5	27,5 ±30,6	19,7 ±27,1
Ich bin in Remission von Krebs / Ich bin ein Krebsüberlebender (n=1,965)	34,6 ±30,8	12,9 ±23,5	15,4 ±24,0	12,0 ±22,3
Analysepopulation (n=2,643)	37,5 ±31,8	17,6 ±27,5	19,2 ±27,3	14,6 ±24,8
Referenz (n=23,553)*	28,9 ±31,9	21,1 ±31,3	17,5 ±28,4	9,0 ±20,3

*Daten zeigen Mittelwert±SD; FA, Fatigue (Müdigkeit); NV, Übelkeit und Erbrechen; PA, Schmerzen; DY, Atemnot; SL, Schlaflosigkeit; AP, Appetitverlust; CO, Verstopfung; DI, Durchfall; * Referenzpopulation hauptsächlich basierend auf Daten aus klinischen Krebsstudien und epidemiologischen Untersuchungen mit insgesamt 23.553 Fällen(47).*

Im Gegensatz dazu lagen die Werte der Symptomskalen der beiden anderen Subgruppen sowie der Gesamtkohorte tendenziell sogar über denen der Referenzpopulation, was auf eine insgesamt höhere symptomatische Belastung in der Analysepopulation hinweist (siehe Tabelle 21).

4 DISKUSSION

Ziel der vorliegenden Arbeit war es zu untersuchen, ob der von der EORTC Quality of Life Gruppe neu entwickelte Fragebogen QLQ-F17 als gekürzte Version des QLQ-C30 als neuer Kernfragebogen analog zum QLQ-C30 verwendet werden kann. Diese Untersuchung war insbesondere deshalb erforderlich, da der QLQ-F17 bislang ausschließlich auf der Website der EORTC veröffentlicht wurde, ohne dass formale Publikationen zu seiner Validierung oder psychometrischen Prüfung vorlagen. Entsprechend fehlte bislang auch die empirische Evidenz dafür, dass die Funktionsskalen beider Fragebögen tatsächlich vergleichbare Ergebnisse liefern. Hintergrund der Notwendigkeit dieser methodischen Prüfung ist die strukturelle Veränderung des QLQ-F17 im Vergleich zum QLQ-C30. Die entfernten Items zur Abfrage der Symptome befinden sich in der Mitte des QLQ-C30, sodass überprüft werden musste, ob deren Wegfall das Antwortverhalten auf die nachfolgenden Fragen beeinflusst. So konnten bereits verschiedene Studien zeigen, dass der Kontext und die Position von Fragen das Antwortverhalten, z.B. durch sogenannte Reihenfolgeeffekte systematisch beeinflussen können (19),(20). Insbesondere der Übergang von symptombezogenen zu funktionsbezogenen Items im QLQ-C30 könnte die Wahrnehmung und Bewertung von funktionalen Fragen beeinflussen. Durch das Weglassen dieser symptombezogenen Fragen im QLQ-F17 ändert sich somit nicht nur die Länge, sondern auch der inhaltliche Fluss des Fragebogens. Vor diesem Hintergrund wurde geprüft, ob diese strukturelle Änderung zu klinisch relevanten Unterschieden in den identischen Items und den daraus resultierenden Skalenwerten führt. Nur durch Nachweis der Äquivalenz kann der QLQ-F17 als neuer Kernfragebogen der EORTC den QLQ-C30 zur Abfrage der funktionsbezogenen Lebensqualität und des allgemeinen Wohlbefindens in der klinischen Forschung ersetzen.

Zur Beantwortung dieser Frage wurde eine randomisierte Studie im Cross-over Design durchgeführt. Die Studie war als internationale Studie konzipiert, sodass Patienten aus insgesamt elf verschiedenen Ländern eingeschlossen werden konnten, was eine hohe kulturelle Vielfalt gewährleistete. Im Rahmen des Cross-over-Designs füllten die teilnehmenden Patienten entweder zuerst den QLQ-F17 oder den QLQ-C30 aus. Nach einer kurzen Wash-out-Phase, in der allgemeine Fragen zum Patienten, zum Krankheitsstatus, zur körperlichen Aktivität sowie zu weiteren Aspekten des Alltags

gestellt wurden, sollte der jeweils anderen Fragebogen beantwortet werden. Dieses Studiendesign ermöglichte einerseits einen direkten Vergleich der beiden Fragebögen zwischen randomisierten Gruppen der ersten Erhebung und andererseits gepaarte Vergleiche innerhalb desselben Patienten.

Zur Beurteilung der Relevanz der Effekte aus den multiplen Regressionsanalysen wurde im Vorfeld ein Äquivalenzintervall definiert, innerhalb dessen Unterschiede zwischen den beiden Fragebögen innerhalb einer Skala als klinisch nicht relevant angesehen wurden. In diesem Fall wurden die Fragebögen bzw. die entsprechenden Skalen als äquivalent betrachtet. Die Wahl eines einheitlichen Äquivalenzintervalls von]–5; 5[für alle Skalen wurde bereits vor der Datenerhebung im Rahmen der Studienplanung festgelegt. Grundlage für die Festlegung dieses Intervalls war eine initiale, umfassende Literaturrecherche zum Thema der kleinsten klinisch relevanten Unterschiede (MCID, minimal clinically important difference) des QLQ-C30 über verschiedene Entitäten und klinische Settings hinweg. Primär basierte die Definition des Intervalls auf der Übersichtsarbeit von Musoro et al., die für sämtliche QLQ-C30-Skalen 21 klinische Studien zum Thema MCID bei unterschiedlichen Tumorentitäten und Studientypen systematisch zusammenfasst (44). Diese Arbeit liefert einerseits eine fundierte Grundlage für die Planung klinischer Studien so zum Beispiel bei der Auswahl spezifischer Skalen als primäre Endpunkte. Andererseits zeigt sie deutlich, dass nahezu alle publizierten MCID-Werte ≥ 5 liegen, was die Rationale für die Wahl eines für alle Skalen gemeinsamen und konservativen Äquivalenzintervalls von ±5 Punkten war. Zusätzlich zu diesen Überlegungen wurde unter Berücksichtigung der Empfehlung der EMA Guideline zur Wahl von nicht-Unterlegenheitsschranken auch ein statistisches Advisory Board in die finale Definition des Äquivalenzintervalls mit einbezogen, das die gewählten Grenzen von –5 bis +5 inhaltlich wie methodisch bestätigte (24).

4.1 BEWERTUNG DER ÄQUIVALENZ

Das Hauptergebnis der vorliegenden Studie ist der statistische Nachweis, dass die Funktionsskalen sowie die Skala zum globalen Gesundheitszustand/Lebensqualität des QLQ-F17 äquivalente Ergebnisse zu denen des QLQ-C30 liefern. Diese Äquivalenz konnte sowohl in Zwischen-Gruppenvergleichen als auch in gepaarten Analysen nachgewiesen werden.

In den Zwischen-Gruppenvergleichen zeigten sich auf Item-Ebene innerhalb der DIF-Analysen erste leichte Abweichungen im Antwortverhalten ab Item 8 des QLQ-F17, welches im QLQ-C30 als Item 20 direkt im Anschluss an die Symptomfragen positioniert ist. Weitere kleinere Verschiebungen wurden für die Items 9, 10, 14 und 15 des QLQ-F17 identifiziert. Die statistisch signifikanten DIF-Effekte waren jedoch durchweg sehr gering, mit R^2-Werten von unter 0,01. Die statistische Signifikanz trotz geringer Effektstärke ist auf die hohe Fallzahl zurückzuführen, bei der auch minimale Unterschiede statistisch signifikant werden können. Somit konnten im Rahmen dieser Analysen keine relevanten Unterschiede im Antwortverhalten der einzelnen Fragen zwischen den beiden Fragebögen identifiziert werden und auf Äquivalenz geschlossen werden. Trotz der geringen Größe der Effekte unterstreicht dieses erste Ergebnis die Relevanz der Fragestellung und zeigt, dass Reihenfolgeneffekte, also der Einfluss der Position von Fragen auf die Beantwortung der nachfolgenden Fragen tatsächlich vorhanden sind.

Konsistent dazu zeigten auch die die multiplen Regressionsanalysen der Zwischen-Gruppenvergleichen bei den Vergleichen der Skalenwerte analoge Muster. Während für die ersten sieben Items mit Hilfe der DIF-Analysen keine Unterschiede festgestellt wurden und sich somit auch nahezu keine Unterschiede für die Funktionsskalen „Physische Funktion“ und „Rollenfunktion“ zeigten, konnten bei den Skalen „Emotionale Funktion“, „Kognitive Funktion“, „Soziale Funktion“ sowie bei der Skala zum „Globalen Gesundheitszustand/Lebensqualität“ Abweichungen zwischen den beiden Fragebögen identifiziert werden. Insgesamt waren aber alle geschätzten Mittelwertsdifferenzen der Skalen inkl. der zugehörigen 95%-Konfidenzintervalle innerhalb des vordefinierten Äquivalenzbereichs, so dass auch diese Analysen den statistischen Nachweis der Äquivalenz erbringen konnten.

Neben den Zwischen-Gruppenvergleichen wurde die Äquivalenz auch durch die gepaarten Messungen (Innerhalb-Gruppenvergleiche) beider Fragebögen innerhalb jedes Patienten unter Verwendung linearer Modelle untersucht. Dabei zeigte sich bei den Effekten ein nahezu identisches Muster wie bei den Zwischen-Gruppenvergleichen, sowohl in der Richtung der Effekte als auch in deren Relation zueinander. Der entscheidende Unterschied bestand jedoch darin, dass die Differenzen bei den gepaarten Analysen deutlich geringer ausfielen, näher an der Null

lagen und somit weiter von den festgelegten Äquivalenzgrenzen entfernt waren. Zudem waren die Konfidenzintervalle deutlich schmäler, was bei gepaarten Stichproben zu erwarten war. Auch weitere Analysen innerhalb der gepaarten Vergleiche ergaben keine auffälligen Abweichungen zwischen den Skalen „Körperliche Funktion“ bzw. „Rollenfunktion“ (Items 1–7) und den übrigen Funktionsskalen. Somit konnte auch im Rahmen der Zwischen-Gruppenvergleiche die statistische Äquivalenz gezeigt werden.

Obwohl in den statistischen Modellen der Zwischen-Gruppenvergleiche keine signifikante Interaktion zwischen der Fragebogenversion und dem Erhebungszeitpunkt festgestellt wurde, zeigten die Mittelwertvergleiche, dass die Unterschiede zwischen den Skalen bei der ersten Erhebung tendenziell größer waren als bei der zweiten Erhebung. Dieser Hinweis auf einen möglichen Carry-over-Effekt, welcher vermutlich durch die vergleichsweise kurze Wash-out-Phase bedingt war, legt nahe, die Ergebnisse der Innerhalb-Gruppenvergleiche vorsichtig zu interpretieren und primär als Sensitivitätsanalysen zu den Zwischen-Gruppenvergleichen zu verstehen.

Insgesamt über alle Modelle und Analysen betrachtet zeigen die Ergebnisse, dass trotz der nachgewiesenen Äquivalenz der Fragebögen dennoch die im QLQ-C30 enthaltenen Symptomfragen einen zwar sehr kleinen aber dennoch direkten oder auch indirekten Einfluss auf die Beantwortung der nachfolgenden Funktionsskalen haben. Dies hat sich primär durch deren Wegfall im QLQ-F17 in leicht veränderten Antwortmustern gezeigt. Betrachtet man die Richtungen der identifizierten Effekte, so lassen sich folgende Schlussfolgerungen ziehen.

Bei Frage 8 des QLQ-F17 bzw. Frage 20 des QLQ-C30 („Hatten Sie Schwierigkeiten, sich auf etwas zu konzentrieren, z. B. auf Zeitunglesen oder Fernsehen?“) zeigte sich für den QLQ-F17 eine höhere Wahrscheinlichkeit besserer Funktionalität im Vergleich zum QLQ-C30. Dies deutet darauf hin, dass sowohl die Symptomfragen im QLQ-C30 einen negativen Einfluss auf diese Frage haben könnten als auch dass Frage 7 im QLQ-F17 einen positiven Einfluss auf das Antwortverhalten ausübt.

Ein genau entgegengesetztes Bild zeigte sich bei den Fragen 9 („Fühlten Sie sich angespannt?“), 10 („Haben Sie sich Sorgen gemacht?“) und 15 („Hat Ihr körperlicher Zustand oder Ihre medizinische Behandlung Ihr Zusammensein oder Ihre

gemeinsamen Unternehmungen mit anderen Menschen beeinträchtigt?“) des QLQ-F17 bzw. den entsprechenden Fragen 21, 22 und 27 im QLQ-C30. Hier ergab sich in den DIF-Analysen ein leichter Shift nach links, was bedeutet, dass diese Fragen im QLQ-F17 tendenziell mit schlechterer Funktionalität beantwortet wurden als im QLQ-C30. Eine inhaltliche Begründung lässt sich auf Basis der vorliegenden Analysen nicht abschließend geben. Möglich ist ein negativer Einfluss der ersten sieben Fragen im QLQ-F17 auf das Antwortverhalten dieser Items. So könnten etwa eine eingeschränkte körperliche Funktion oder reduzierte Rollenfunktion das Gefühl von Anspannung und Sorgen verstärken. Ebenso könnte ein schlechter physischer Zustand zu einer negativen Bewertung sozialer Aspekte wie Zusammensein oder gemeinsame Aktivitäten führen.

Bei Frage 14 des QLQ-F17 („Hat Ihr körperlicher Zustand oder Ihre medizinische Behandlung Ihr Familienleben beeinträchtigt?“) bzw. Frage 26 des QLQ-C30 zeigte sich eine nicht-uniforme DIF, bei der die Beantwortung im QLQ-F17 tendenziell schlechter ausfiel, wenn die zugrundeliegende Lebensqualität hoch war (kleinere Theta Werte).

Zur weiteren Untersuchung der Effekte der DIF-Items wurden zusätzlich Subgruppenanalysen nach Geschlecht und Symptomlast durchgeführt. So zeigten sich zwischen Männern und Frauen insgesamt nur geringe Unterschiede, was auf eine weitgehende Geschlechterunabhängigkeit der Reihenfolgeeffekte hinweist. Auffällig war jedoch, dass bei Männern kein DIF bei den beiden Items zur sozialen Funktion identifiziert wurde. Dies könnte auf geschlechterspezifische Unterschiede in der Stabilität der Bewertung sozialer Funktion hindeuten, etwa dahingehend, dass Männer diese weniger stark im Kontext anderer Items (z. B. Symptome oder emotionale Aspekte) bewerten. Eine abschließende Interpretation bedarf jedoch weiterer Untersuchungen in diesem Bereich.

Deutlich differenzierter zeigten sich die Ergebnisse in Abhängigkeit von der Symptomlast. Besonders Frage 8 des QLQ-F17, welche im QLQ-C30 unmittelbar auf die Symptomskalen folgt, zeigte unterschiedliche Verhaltensmuster in Abhängigkeit vom Ausmaß der Symptombelastung. Während bei Patienten mit niedriger Symptomlast kein DIF-Effekt auftrat, nahm die Effektstärke mit steigender Symptomlast kontinuierlich zu (mittlere Symptome: $Pseudo - \mathrm{R}^2_{\text{Nagelkerke}}$=0,012; hohe

Symptome: $\mathrm{Pseudo} - \mathrm{R}^2_{\mathrm{Nagelkerke}}$=0,042). Dieses Muster deutet auf einen möglichen systematischen Zusammenhang zwischen Symptombelastung und der Bewertung funktionaler Items hin. Je höher die Belastung durch Symptome der Behandlung, desto negativer fällt die Bewertung nachfolgender Fragen aus. Zwar bleiben die Effektstärken formal im kleinen Bereich, ihre klinische Relevanz sollte jedoch nicht unterschätzt werden, da sie auf kleine, aber systematische Verzerrungen im Antwortverhalten zwischen dem QLQ-C30 und dem QLQ-F17 hinweisen könnten.

Auch bei den weiteren DIF-Items zeigte sich bei mittlerer Symptomlast im Vergleich zu niedriger Symptomlast ein Anstieg der Effektstärken, was auf vergleichbare Zusammenhänge schließen lässt. Das Fehlen weiterer Effekte in der Gruppe mit hoher Symptomlast dürfte hingegen weniger auf eine tatsächliche Abwesenheit, sondern eher auf die geringe Fallzahl und damit eingeschränkte statistische Power zurückzuführen sein.

Diese Ergebnisse liefern erste empirische Hinweise darauf, dass Symptomlast als Moderator für Reihenfolgeeffekte wirken kann. Zukünftige Studien sollten diese Zusammenhänge gezielt untersuchen, um die Bewertung funktionaler Lebensqualität in Abhängigkeit vom klinischen Zustand noch besser verstehen und methodisch absichern zu können.

4.2 PSYCHOMETRISCHE EIGENSCHAFTEN

Neben der Hauptfragestellung zur Untersuchung der Äquivalenz der beiden Fragebögen QLQ-F17 und QLQ-C30 wurden im Rahmen der Studie auch die psychometrischen Eigenschaften des QLQ-F17 betrachtet. Dabei wurde zunächst eine konfirmatorische Faktorenanalyse durchgeführt, welche die vorgegebene Faktorenstruktur mit 6 Faktoren und der Vorgabe, welche Items auf welche Skalen laden sollen untersuchte. Es zeigte sich, dass alle Items hohe Faktorladungen deutlich über 0,4 aufwiesen. Dies bestätigte die vorgegebene Faktorenstruktur des QLQ-F17 mit 6 Faktoren mit einer sehr guten Item-Faktor Zuordnung. Die Items des QLQ-F17 hängen somit stark mit dem zugehörigen latenten Konstrukt (dem Faktor) zusammen und geben somit auch eine erste Evidenz für die konvergente Validität innerhalb der jeweiligen Skalen. Zur Beurteilung der Modellanpassung wurden weiterhin die 4 Kenngrößen Comparative-Fit-Index (CFI), Tucker-Lewis-Index (TLI), Root-Mean-

Square-Error of Approximation (RMSEA) und Standardized Root Mean Square Residual (SRMR) berechnet. Dabei zeigten sowohl der CFI mit 0.941, der RMSEA mit 0,079 als auch das SRMR mit 0,050 Werte, die nach Byrne und Kline für eine gute Modellanpassung sprechen (61),(62). Der TLI von 0,922 liegt zwar knapp unter der von Byrne vorgeschlagenen Grenze von 0,95 aber deutlich über den von Kline benannten Grenze von 0,85. Somit zeigt der F17 insgesamt eine gute Modellanpassung, die auch vergleichbar mit den Werten ist, die zum QLQ-C30 publiziert wurden. Die interne Konsistenz der Skalen (Reliabilität) wurde durch die Berechnung von Cronbachs α beurteilt. Dabei zeigte sich für alle Skalen des QLQ-F17 ein Wert über dem empfohlenen Wert von 0,7 was für eine gute Reliabilität des QLQ-F17 spricht. Als letzte Beurteilung der psychometrischen Eigenschaften des QLQ-F17 wurde die Konstruktvalidität untersucht. Dabei zeigte sich für fast alle Items eine gute Trennschärfe im Sinne einer guten Korrelation von über 0,4 mit der eigenen Skala im Gegensatz zu den Korrelationen zu den Fremdskalen. In nur 3 Fällen konnte bei einem Item ein Skalierungsfehler, sprich eine höhere Korrelation zu einer Fremdskala als zur eigenen Skala festgestellt werden. So zeigte Frage 5 „Brauchen Sie Hilfe beim Essen, Anziehen, Waschen oder Benutzen der Toilette?“ mit einer Korrelation von 0,49 auf die Kognitive Funktion und 0,50 auf die Rollenfunktion gleich zweimal eine höhere Korrelation im Vergleich zur Korrelation zur eigenen Skala der Körperlichen Funktion mit einem Wert von 0,47. Ebenso zeigte sich bei Frage 8 „Hatten Sie Schwierigkeiten, sich auf etwas zu konzentrieren, z. B. auf Zeitunglesen oder Fernsehen?“ eine um 0,01 höhere Korrelation zur Fremdskala der emotionalen Funktion mit 0,59 im Vergleich zur Korrelation mit der eigenen Skala, der kognitiven Funktion von 0,58. Diese sehr kleinen Skalierungsfehler fallen jedoch nicht ins Gewicht und können vernachlässigt werden.

4.3 Repräsentativität der Ergebnisse

Die Hauptergebnisse der Studie basieren auf einer sehr großen und internationalen Stichprobe mit insgesamt 2.643 Patienten. Da es im Rahmen der Datenerhebung keine Ein- und Ausschlusskriterien außer einer vorhandenen oder zurückliegenden Krebserkrankung gab ist das Patientenkollektiv als sehr heterogen einzustufen. Dies zeigte sich z.B. in der Spannweite des Alters mit Patienten im Alter zwischen 18 und 92 oder auch in der großen Vielfalt an unterschiedlichen Krebsentitäten. Zudem

wurden Patienten aus elf verschiedenen Ländern und sechs Sprachregionen eingeschlossen, was insgesamt zu einer hohen Verallgemeinbarkeit der Ergebnisse führt. Bei einem Vergleich der Stichprobe mit der Referenzpopulation aus dem EORTC Reference Value Manual konnte zudem gezeigt werden, dass sich die Mittelwerte und Standardabweichungen der Funktionsskalen kaum unterscheiden und nahezu identisch sind (47). Trotz der Tatsache, dass in der Studienstichprobe fast dreiviertel der Patienten in Remission sind bzw. den Krebs besiegt haben, zeigte sich bei den Symptomskalen kaum ein Unterschied zur Referenzpopulation. Es konnte hingegen gezeigt werden, dass die Patienten der vorliegenden Studie in einzelnen Subgruppen sogar höhere körperliche Beeinträchtigungen berichten als in der Referenzpopulation. So zeigten Patienten in der Altersgruppe unter 50 Jahre der Studienpopulation deutlich niedrigere mittlere Scores in den Funktionsskalen als in der Referenzpopulation. Eine Erklärung hierfür könnte der deutlich höhere Anteil mit 42% an neu diagnostizierten Patienten bzw. an Patienten, die sich in einer laufenden Therapie befinden im Vergleich zu den restlichen Altersgruppen sein. Hier liegt der Anteil deutlich darunter mit 24% in der Altersgruppe von 50-59 Jahren, 17% in der Altersgruppe von 60-69 Jahren und 18% in der Altersgruppe der über 70-jährigen.

4.4 LIMITATIONEN

Im Rahmen der vorliegenden Arbeit konnten zwei Limitationen identifiziert werden, welche im Folgenden beschrieben und diskutiert werden.

4.4.1 DATENERHEBUNG

Eine mögliche Limitation im Rahmen der Studie ist die Art der Datenerhebung. Im Gegensatz zu der üblichen Herangehensweise der Datenerhebung im direkten Patientenkontakt in den Kliniken, wurde die Datenerhebung für diese Studie von einem Unternehmen durchgeführt. Hierfür wurde die Firma Kantar engagiert, ein professionelles, zertifiziertes und weltweit führendes Unternehmen für Marketingdaten und Panelanalysen. Für den Bereich im Gesundheitswesen hält Kantar eine Vielzahl an verschiedenen Patientenpanels vor, die primär für Studien der Pharmazeutischen Industrie verwendet werden. Somit wurden die Daten ohne direkten Patientenkontakt erhoben und die Angaben zur Soziodemografie sowie zur Krebserkrankung an sich können nicht direkt überprüft werden. Mittels der Überprüfung der Datenqualität durch

die eingeführten Qualitätschecks konnte jedoch gezeigt werden, dass nahezu alle Patienten die Fragebögen plausibel und in einem angemessen Zeitrahmen ausgefüllt haben. Ebenso ist durch die gute Übereinstimmung der Mittelwerte und Standardabweichungen der Skalen des QLQ-C30 zu den von der EORTC publizierten Referenzpopulation davon auszugehen, dass es sich um ein repräsentatives Kollektiv von Krebspatienten handelt. Der geringe Anteil an Patienten mit inkonsistenten Antworten oder zu kurzen Antwortzeiten der Fragen ist so auch bei einer klassischen Datenerhebung zu erwarten. Die Vorteile der Datenerhebung in einem kurzen Zeitraum von nur knapp 2 Monaten von über 2.500 Patienten überwiegt somit den möglichen Nachteilen, wenn man bedenkt, dass eine manuelle Erhebung von über 2.500 Patienten kaum unter 10 Jahren schaffbar gewesen wäre. Insgesamt betrachtet ist diese Art der Datenakquise in der heutigen Zeit eine vertretbare und valide Methode um relativ große Stichproben in kurzen Zeiträumen zu erheben (z.B. Nolte et al. (26)).

4.4.2 Cross-over Design und Wash-out Phase

Eine weitere Limitation der Arbeit bezieht sich auf das Cross-over Design und die gewählte Wash-out Phase. Die Studie wurde so konzipiert, dass die Patienten beide Fragebögen innerhalb einer Session beantworten sollten, nur mit einer kurzen Wash-out Phase von ca. 10 Minuten zwischen der Erhebung der beiden Fragebögen QLQ-C30 und QLQ-F17. Die Gründe für diese Entscheidung waren einerseits die deutlich erhöhten Kosten für die Durchführung der Patientenbefragung, wenn die Patienten eine zweite Befragung an einem anderen Tag hätten durchführen sollen. Zum anderen sollte das Risiko, dass ein Anteil an Patienten die zweite Befragung nicht mehr durchführt möglichst geringgehalten werden, um keine Verzerrung der Ergebnisse durch Studienabbrecher (sogenannter Attrition Bias) zu erhalten. Schaut man in die Literatur, so besteht eine große Heterogenität in der Wahl der Dauer der Wash-out Phasen bei Test-Retest Erhebungen für die Evaluation von Fragebögen. So finden sich in der Literatur Studien mit kurzen Wash-out Phasen von 3 Stunden bis hin zu Studien mit Wash-out Phasen von bis zu 2 Wochen. Dabei gibt es in der Wissenschaft keinen Konsens welches Zeitintervall am sinnvollsten ist. So konnte z.B. von Marx et al. in einer Studie zum direkten Vergleich zweier Zeitintervalle (2 Tage und 2 Wochen) kein Unterschied in der Test-Retest Reliabilität nachgewiesen werden (67). Der entscheidende Aspekt in dieser Frage ist die mögliche Beeinflussung der Zeit der

Wash-out Phase auf die Beantwortung des zweiten Fragebogens. So ist bei einer zu lange gewählten Wash-out Phase das Risiko gegeben, dass zwischen den beiden Erhebungen dem Patienten etwas widerfährt, was einen negativen oder positiven Einfluss auf die zweite Messung haben kann. Somit kann in diesem Fall keine Konsistenz und somit keine Vergleichbarkeit zwischen den beiden Messungen erwartet werden. Ist die Wash-out Phase hingegen zu kurz gewählt, so ist es möglich, dass sich die Patienten an ihre in der ersten Erhebung gegebenen Antworten erinnern und diese ohne erneutes Nachdenken identisch wiedergeben können. Obwohl die Wash-out Phase in der vorliegenden Studie von nur 10 Minuten sehr kurz erscheint, zeigte sich jedoch, dass trotz des kurzen Zeitabstands die Antworten auf dieselben Fragen oft nicht identisch waren, sondern einige Abweichungen vom tatsächlichen Ergebnis aufwiesen. Trotzdem konnte in weiteren Analysen gezeigt werden, dass Carry-over-Effekte nicht auszuschließen sind, so dass die primäre Beurteilung der Äquivalenz auf dem randomisierten Design basiert, was nach aktuellen wissenschaftlichen Standards die strengste Vergleichsmethode darstellt.

4.5 PERSPEKTIVEN

Die vorliegenden Ergebnisse sprechen für eine zukünftige Anwendung des QLQ-F17 als Kerninstrument in Kombination mit passenden Modulen und Einzelfragen zu Symptomen aus der EORTC Item Library in onkologischen klinischen Studien. Durch den Nachweis der Äquivalenz bleiben Ergebnisse aus zukünftigen Studien mit dem QLQ-F17 weiterhin vergleichbar mit bereits publizierten Studienergebnissen auf Basis des QLQ-C30. Ein nächster Schritt in der Weiterentwicklung des QLQ-F17 betrifft die Frage, ob sich aus den funktionalen Skalen ein valider Summenscore ableiten lässt, der die patientenberichtete Funktionsfähigkeit auf einer Gesamtskala abbildet. Dies könnte sowohl die Interpretation der funktionalen Lebensqualität in klinischen Studien erleichtern, als auch die Vergleichbarkeit zwischen verschiedenen Studiengruppen verbessern. Darüber hinaus wird es notwendig sein, Referenz- bzw. Normwerte für den QLQ-F17 zu etablieren, beispielsweise durch repräsentative Erhebungen in der Normalbevölkerung als auch bei Krebspatienten, idealerweise in unterschiedlichen Ländern und verschiedenen Sprachen. Solche Vergleichswerte ermöglichen eine differenziertere Einordnung individueller Lebensqualitätswerte und deren Veränderungen über die Zeit hinweg sowie eine genauere Bewertung von

Studienergebnissen einzelner Patientenkohorten. Perspektivisch stellt auch ein systematischer Vergleich der funktionalen Lebensqualität zwischen onkologischen und nicht-onkologischen Patientengruppen einen interessanten Ansatz dar. Dies könnte helfen, krankheitsspezifische Belastungsmuster besser zu verstehen und die potenzielle Anwendbarkeit des QLQ-F17 auch über die Onkologie hinaus zu prüfen.

5 SCHLUSSFOLGERUNG

Der QLQ-F17 erweist sich als ein zuverlässiges Instrument zur Erfassung der funktionalen Lebensqualität und des globalen Gesundheitszustands von Krebspatienten. Die vorliegenden Ergebnisse bestätigen die inhaltliche Äquivalenz der funktionalen Skalen des QLQ-F17 im Vergleich zum etablierten QLQ-C30. Damit bietet der QLQ-F17 die Möglichkeit für eine flexiblere und zugleich zeiteffizientere Erhebung der Lebensqualität, ohne dabei an Aussagekraft und Vergleichbarkeit mit bisherigen Studien verzichten zu müssen.

Zwar konnten auf Einzelebene kleinere Kontexteffekte identifiziert werden, diese sind jedoch aus klinischer Sicht vernachlässigbar. Der QLQ-F17 ermöglicht es somit, in Studien, die sich auf behandlungsspezifische Symptome aus der EORTC Item Library fokussieren ein kompaktes Kerninstrument zur Erfassung funktionaler Lebensqualität einzusetzen. Durch diese Flexibilität in der Anwendung trägt der QLQ-F17 dazu bei, patientenberichtete Endpunkte einfacher und ressourcenschonender in onkologische Studien und die klinische Praxis zu integrieren.

6 LITERATURVERZEICHNIS

1 Zeman F, Giesinger JM, Pukrop T, et al. The EORTC QLQ-F17 as a shortened version of the EORTC QLQ-C30 to assess self-reported functioning in cancer patients: investigating equivalence and psychometric properties in a randomized cross-over trial. *EClinicalMedicine* 2025; 84: 103262. https://doi.org/10.1016/j.eclinm.2025.103262

2 Germann P. The Quality of Life Turn: The Measurement and Politics of Well-Being in the 1970s. *KNOW: A Journal on the Formation of Knowledge* 2020; 4: 295–324. https://doi.org/10.1086/710511

3 Montazeri A. Quality of life data as prognostic indicators of survival in cancer patients: an overview of the literature from 1982 to 2008. *Health Qual Life Outcomes* 2009; 7: 102. https://doi.org/10.1186/1477-7525-7-102

4 Constitution of the World Health Organization. 1946. *Bull World Health Organ* 2002; 80: 983–84.

5 Calman KC. Quality of life in cancer patients--an hypothesis. *J Med Ethics* 1984; 10: 124–27. https://doi.org/10.1136/jme.10.3.124

6 World Health Organization. Programme on mental health: WHOQOL user manual. Technical documents. World Health Organization, 1998.

7 Deshpande PR, Rajan S, Sudeepthi BL, Abdul Nazir CP. Patient-reported outcomes: A new era in clinical research. *Perspect Clin Res* 2011; 2: 137–44. https://doi.org/10.4103/2229-3485.86879

8 Carr AJ, Higginson IJ. Are quality of life measures patient centred? *BMJ* 2001; 322: 1357–60. https://doi.org/10.1136/bmj.322.7298.1357

9 Castro MG, Wang MC. Quality of Life and Patient-Centered Outcomes. In: Daaleman TP, Helton MR, eds. Chronic Illness Care. Cham: Springer International Publishing, 2023: 511–24.

10 Allison PJ, Locker D, Feine JS. Quality of life: a dynamic construct. *Soc Sci Med* 1997; 45: 221–30. https://doi.org/10.1016/s0277-9536(96)00339-5

11 McDowell I. Measuring Health. A Guide to Rating Scales and Questionnaires. New York: Oxford University Press USA - OSO, 2006.

12 Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in

international clinical trials in oncology. *J Natl Cancer Inst* 1993; 85: 365–76. https://doi.org/10.1093/jnci/85.5.365

13 Cocks K, Wells JR, Johnson C, et al. Content validity of the EORTC quality of life questionnaire QLQ-C30 for use in cancer. *Eur J Cancer* 2023; 178: 128–38. https://doi.org/10.1016/j.ejca.2022.10.026

14 Fayers P, Bottomley A. Quality of life research within the EORTC-the EORTC QLQ-C30. European Organisation for Research and Treatment of Cancer. *Eur J Cancer* 2002; 38 Suppl 4: S125-33. https://doi.org/10.1016/s0959-8049(01)00448-8

15 Aaronson NK, Cull A, Kaasa S, Sprangers MA. The EORTC Modular Approach to Quality of Life Assessment in Oncology. *International Journal of Mental Health* 1994; 23: 75–96. https://doi.org/10.1080/00207411.1994.11449284

16 Kulis D, Bottomley A, Whittaker C, et al. The Use of The Eortc Item Library To Supplement Eortc Quality of Life Instruments. *Value in Health* 2017; 20: A775. https://doi.org/10.1016/j.jval.2017.08.2236

17 Piccinin C, Kulis D, Bottomley A, et al. EORTC quality of life group item library user guidelines. Brussels: EORTC, 2022.

18 U.S. Department of Health and Human Services, Food and Drug Administration. Core Patient-Reported Outcomes in Cancer Clinical Trials: Guidance for Industry; Available from: https://www.fda.gov/media/149994/download, 2024.

19 Wänke M, Schwarz N. Reducing Question Order Effects: The Operation of Buffer Items. In: Lyberg L, ed. Survey measurement and process quality. New York: Wiley, 2010: 115–40.

20 Siminski P. Order Effects in Batteries of Questions. *Qual Quant* 2008; 42: 477–90. https://doi.org/10.1007/s11135-006-9054-2

21 Schwarz N, Strack F, Mai H-P. Assimilation and Contrast Effects in Part-Whole Question Sequences: A Conversational Logic Analysis. *Public Opinion Quarterly* 1991; 55: 3. https://doi.org/10.1086/269239

22 Lesaffre E. Superiority, equivalence, and non-inferiority trials. *Bull NYU Hosp Jt Dis* 2008; 66: 150–54.

23 U.S. Department of Health and Human Services, Food and Drug Administration. Non-inferiority clinical trials to establish effectiveness; Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/non-inferiority-clinical-trials, 2016.

24 Committee for Medicinal Products for Human Use, the European Medicines Agency. Guideline on the choice of the non-inferiority margin; Available from: https://www.ema.europa.eu/en/choice-non-inferiority-margin-scientific-guideline, 2005.
25 Feak C, Swales J. Telling a Research Story: Writing a Literature Review. Ann Arbor, MI: University of Michigan Press/ELT, 2009.
26 Nolte S, Coon C, Hudgens S, Verdam MGE. Psychometric evaluation of the PROMIS® Depression Item Bank: an illustration of classical test theory methods. *J Patient Rep Outcomes* 2019; 3: 46. https://doi.org/10.1186/s41687-019-0127-0
27 Eurofound, Dubois H, Jungblut J-M, et al. European quality of life survey 2016 – Quality of life, quality of public services, and quality of society – Overview report. Publications Office, 2017.
28 Kuliś D, Holzner B, Koller M, et al. Guidance on the Implementation and Management of EORTC Quality of Life Instruments in Electronic Applications; Available from: https://qol.eortc.org/manuals, 2018.
29 Kulis D, Bottomley A, Velikova G. Translation procedure Manual.; Available from: https://qol.eortc.org/manuals. Brussels: European Organization for Research and Treatment of Cancer, 2017.
30 International Conference on Harmonisation. E9: Guidance on statistical principles for clinical trials. September 16, 1998: Federal Register 63(179).
31 Vet HCW de, Mokkink LB, Terwee CB. Minimal Clinically Important Difference (MCID). In: Michalos AC, ed. Encyclopedia of Quality of Life and Well-Being Research. Dordrecht: Springer Netherlands, 2014: 4071–72.
32 Osoba D, Rodrigues G, Myles J, Zee B, Pater J. Interpreting the significance of changes in health-related quality-of-life scores. *J Clin Oncol* 1998; 16: 139–44. https://doi.org/10.1200/JCO.1998.16.1.139
33 Cocks K, King MT, Velikova G, et al. Evidence-based guidelines for interpreting change scores for the European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30. *Eur J Cancer* 2012; 48: 1713–21. https://doi.org/10.1016/j.ejca.2012.02.059
34 Dirven L, Musoro JZ, Coens C, et al. Establishing anchor-based minimally important differences for the EORTC QLQ-C30 in glioma patients. *Neuro Oncol* 2021; 23: 1327–36. https://doi.org/10.1093/neuonc/noab037

35 Gamper EM, Musoro JZ, Coens C, et al. Minimally important differences for the EORTC QLQ-C30 in prostate cancer clinical trials. *BMC Cancer* 2021; 21: 1083. https://doi.org/10.1186/s12885-021-08609-7

36 Kawahara T, Taira N, Shiroiwa T, et al. Minimal important differences of EORTC QLQ-C30 for metastatic breast cancer patients: Results from a randomized clinical trial. *Qual Life Res* 2022; 31: 1829–36. https://doi.org/10.1007/s11136-021-03074-y

37 Koller M, Musoro JZ, Tomaszewski K, et al. Minimally important differences of EORTC QLQ-C30 scales in patients with lung cancer or malignant pleural mesothelioma - Interpretation guidance derived from two randomized EORTC trials. *Lung Cancer* 2022; 167: 65–72. https://doi.org/10.1016/j.lungcan.2022.03.018

38 Musoro JZ, Sodergren SC, Coens C, et al. Minimally important differences for interpreting the EORTC QLQ-C30 in patients with advanced colorectal cancer treated with chemotherapy. *Colorectal Dis* 2020; 22: 2278–87. https://doi.org/10.1111/codi.15295

39 Musoro JZ, Coens C, Fiteni F, et al. Minimally Important Differences for Interpreting EORTC QLQ-C30 Scores in Patients With Advanced Breast Cancer. *JNCI Cancer Spectr* 2019; 3: pkz037. https://doi.org/10.1093/jncics/pkz037

40 Musoro JZ, Coens C, Greimel E, et al. Minimally important differences for interpreting European Organisation for Research and Treatment of Cancer (EORTC) Quality of life Questionnaire core 30 scores in patients with ovarian cancer. *Gynecol Oncol* 2020; 159: 515–21. https://doi.org/10.1016/j.ygyno.2020.09.007

41 Musoro JZ, Coens C, Singer S, et al. Minimally important differences for interpreting European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire Core 30 scores in patients with head and neck cancer. *Head Neck* 2020; 42: 3141–52. https://doi.org/10.1002/hed.26363

42 Musoro JZ, Bottomley A, Coens C, et al. Interpreting European Organisation for Research and Treatment for Cancer Quality of life Questionnaire core 30 scores as minimally importantly different for patients with malignant melanoma. *Eur J Cancer* 2018; 104: 169–81. https://doi.org/10.1016/j.ejca.2018.09.005

43 Musoro ZJ, Hamel J-F, Ediebah DE, et al. Establishing anchor-based minimally important differences (MID) with the EORTC quality-of-life measures: a meta-

analysis protocol. *BMJ Open* 2018; 8: e019117. https://doi.org/10.1136/bmjopen-2017-019117

44 Musoro JZ, Coens C, Sprangers MAG, et al. Minimally important differences for interpreting EORTC QLQ-C30 change scores over time: A synthesis across 21 clinical trials involving nine different cancer types. *Eur J Cancer* 2023; 188: 171–82. https://doi.org/10.1016/j.ejca.2023.04.027

45 Jaeger SR, Cardello AV. Factors affecting data quality of online questionnaires: Issues and metrics for sensory and consumer research. *Food Quality and Preference* 2022; 102: 104676. https://doi.org/10.1016/j.foodqual.2022.104676

46 Ehrlinger L, Wöß W. A Survey of Data Quality Measurement and Monitoring Tools. *Front Big Data* 2022; 5: 850611. https://doi.org/10.3389/fdata.2022.850611

47 Scott NW, Fayers PM, Aaronson NK, et al. EORTC QLQ-C30 Reference Values Manual. (2nd ed.). Brussels: EORTC Quality of Life Group, 2008.

48 Fayers P, Aaronson N, Bjordal K, et al. EORTC QLQ-C30 Scoring Manual (3rd Edition). *European Organisation for Research and Treatment of Cancer* 2001.

49 Embretson SE, Reise SP. Item Response Theory. New York: Psychology Press, 2013.

50 Scott NW, Fayers PM, Bottomley A, et al. Comparing translations of the EORTC QLQ-C30 using differential item functioning analyses. *Qual Life Res* 2006; 15: 1103-15; discussion 1117-20. https://doi.org/10.1007/s11136-006-0040-x

51 Scott NW, Fayers PM, Aaronson NK, et al. The use of differential item functioning analyses to identify cultural differences in responses to the EORTC QLQ-C30. *Qual Life Res* 2007; 16: 115–29. https://doi.org/10.1007/s11136-006-9120-1

52 Scott NW, Fayers PM, Aaronson NK, et al. Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression. *Health Qual Life Outcomes* 2010; 8: 81. https://doi.org/10.1186/1477-7525-8-81.

53 Crane PK, Gibbons LE, Ocepek-Welikson K, et al. A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. *Qual Life Res* 2007; 16 Suppl 1: 69–84. https://doi.org/10.1007/s11136-007-9185-5

54 Falk RF, Miller NB. A primer for soft modeling. Akron, Ohio: University of Akron Press, 1992.

55 Chambers JM. Graphical methods for data analysis. Boca Raton, FL: CRC Press, 2018.

56 Hoaglin DC, ed. Understanding robust and exploratory data analysis. New York, Weinheim: Wiley, 2000.
57 Schielzeth H, Dingemanse NJ, Nakagawa S, et al. Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods Ecol Evol* 2020; 11: 1141–52. https://doi.org/10.1111/2041-210X.13434
58 Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 1977; 33: 159. https://doi.org/10.2307/2529310
59 Portney LG, Watkins MP. Foundations of clinical research : applications to practice. Philadelphia: F.A. Davis Company, 2015.
60 Hinz A, Singer S, Brähler E. European reference values for the quality of life questionnaire EORTC QLQ-C30: Results of a German investigation and a summarizing analysis of six European general population normative studies. *Acta Oncol* 2014; 53: 958–65. https://doi.org/10.3109/0284186X.2013.879998
61 Kline RB. Principles and practice of structural equation modeling. New York, NY, London: GUILFORD, 2023.
62 Byrne BM. Structural equation modeling with EQS and EQS/Windows. Basic concepts, applications, and programming. Thousand Oaks, Calif.: Sage, 1998.
63 Fayers PM, Machin D. Quality of Life. The assessment, analysis, and reporting of patient-reported outcomes (3rd edition). Chichester, West Sussex, UK, Oxford, UK, Hoboken, NJ: Wiley, 2016.
64 Nunnally JC. Psychometric theory (2nd ed.). New York: McGraw-Hill, 1978.
65 Dwan K, Li T, Altman DG, Elbourne D. CONSORT 2010 statement: extension to randomised crossover trials. *BMJ* 2019; 366: l4378. https://doi.org/10.1136/bmj.l4378
66 Senn SS. Cross-over Trials in Clinical Research. John Wiley & Sons, 2002.
67 Marx RG, Menezes A, Horovitz L, Jones EC, Warren RF. A comparison of two time intervals for test-retest reliability of health status instruments. *J Clin Epidemiol* 2003; 56: 730–35. https://doi.org/10.1016/s0895-4356(03)00084-2

7 ANHANG

A) EORTC QLQ-C30 FRAGEBOGEN IN DER DEUTSCHEN ORIGINALFORM

GERMAN

EORTC QLQ-C30 (Version 3)

Wir sind an einigen Angaben interessiert, die Sie und Ihre Gesundheit betreffen. Bitte beantworten Sie die folgenden Fragen selbst, indem Sie die Zahl einkreisen, die am besten auf Sie zutrifft. Es gibt keine „richtigen" oder „falschen" Antworten. Ihre Angaben werden streng vertraulich behandelt.

Bitte tragen Sie Ihre Initialen ein: ⌊ ⌋⌊ ⌋⌊ ⌋⌊ ⌋

Ihr Geburtsdatum (Tag, Monat, Jahr): ⌊ ⌋⌊ ⌋⌊ ⌋⌊ ⌋⌊ ⌋⌊ ⌋⌊ ⌋⌊ ⌋

Das heutige Datum (Tag, Monat, Jahr): 31 ⌊ ⌋⌊ ⌋⌊ ⌋⌊ ⌋⌊ ⌋⌊ ⌋⌊ ⌋⌊ ⌋

		Überhaupt nicht	Wenig	Ziemlich	Sehr
1.	Bereitet es Ihnen Schwierigkeiten, sich körperlich anzustrengen (z. B. eine schwere Einkaufstasche oder einen Koffer zu tragen)?	1	2	3	4
2.	Bereitet es Ihnen Schwierigkeiten, einen längeren Spaziergang zu machen?	1	2	3	4
3.	Bereitet es Ihnen Schwierigkeiten, eine kurze Strecke außer Haus zu gehen?	1	2	3	4
4.	Müssen Sie tagsüber im Bett liegen oder in einem Sessel sitzen?	1	2	3	4
5.	Brauchen Sie Hilfe beim Essen, Anziehen, Waschen oder Benutzen der Toilette?	1	2	3	4

Während der letzten Woche:

		Überhaupt nicht	Wenig	Ziemlich	Sehr
6.	Waren Sie bei Ihrer Arbeit oder bei anderen tagtäglichen Beschäftigungen eingeschränkt?	1	2	3	4
7.	Waren Sie bei Ihren Hobbys oder anderen Freizeitbeschäftigungen eingeschränkt?	1	2	3	4
8.	Waren Sie kurzatmig?	1	2	3	4
9.	Hatten Sie Schmerzen?	1	2	3	4
10.	Mussten Sie sich ausruhen?	1	2	3	4
11.	Hatten Sie Schlafstörungen?	1	2	3	4
12.	Fühlten Sie sich schwach?	1	2	3	4
13.	Hatten Sie Appetitmangel?	1	2	3	4
14.	War Ihnen übel?	1	2	3	4
15.	Haben Sie erbrochen?	1	2	3	4
16.	Hatten Sie Verstopfung?	1	2	3	4

Bitte wenden

Während der letzten Woche:	**Überhaupt nicht**	**Wenig**	**Ziemlich**	**Sehr**
17. Hatten Sie Durchfall?	1	2	3	4
18. Waren Sie müde?	1	2	3	4
19. Fühlten Sie sich durch Schmerzen in Ihrem alltäglichen Leben beeinträchtigt?	1	2	3	4
20. Hatten Sie Schwierigkeiten, sich auf etwas zu konzentrieren, z. B. auf Zeitunglesen oder Fernsehen?	1	2	3	4
21. Fühlten Sie sich angespannt?	1	2	3	4
22. Haben Sie sich Sorgen gemacht?	1	2	3	4
23. Waren Sie reizbar?	1	2	3	4
24. Fühlten Sie sich niedergeschlagen?	1	2	3	4
25. Hatten Sie Schwierigkeiten, sich an Dinge zu erinnern?	1	2	3	4
26. Hat Ihr körperlicher Zustand oder Ihre medizinische Behandlung Ihr Familienleben beeinträchtigt?	1	2	3	4
27. Hat Ihr körperlicher Zustand oder Ihre medizinische Behandlung Ihr Zusammensein oder Ihre gemeinsamen Unternehmungen mit anderen Menschen beeinträchtigt?	1	2	3	4
28. Hat Ihr körperlicher Zustand oder Ihre medizinische Behandlung für Sie finanzielle Schwierigkeiten mit sich gebracht?	1	2	3	4

Bitte kreisen Sie bei den folgenden Fragen die Zahl zwischen 1 und 7 ein, die am besten auf Sie zutrifft:

29. Wie würden Sie insgesamt Ihren Gesundheitszustand während der letzten Woche einschätzen?

1	2	3	4	5	6	7
sehr schlecht						ausgezeichnet

30. Wie würden Sie insgesamt Ihre Lebensqualität während der letzten Woche einschätzen?

1	2	3	4	5	6	7
sehr schlecht						ausgezeichnet

B) PATIENTENBEFRAGUNG

Umsetzung der Patientenbefragung mit allen Fragen und den jeweiligen Antwortoptionen in der Version A, in welcher der QLQ-C30 zuerst abgefragt wurde. In Version B sind die Fragen des QLQ-C30 und des QLQ-F17 vertauscht. Alle Details zur technischen Umsetzung sind in englischer Sprache und mit rot markiert.

STANDARD-EINFÜHRUNGSBILDSCHIRM – DSGVO-KONFORM – PATIENTENBEFRAGUNG

[SCREEN 1 - PN: TO BE SHOWN TO ALL AND ON THE SAME SCREEN]

Vielen Dank, dass Sie sich bereit erklärt haben, an dieser Studie teilzunehmen.

[PN: THIS SECTION TEXT SHOULD BE COLLAPSIBLE [SEE ARROW IN LEFT HAND SIDE MARGIN]. BY DEFAULT, SHOW COLLAPSED AND ONLY GROUP WHEN ARROW SELECTED]

Die Studie entspricht der DSGVO und allen internationalen/lokalen Datenschutzgesetzen sowie den Richtlinien der Insights Association/EphMRA/BHBIA.

Mit dem Aufrufen des Umfrage-Links erklären Sie sich mit Folgendem einverstanden:

- Ich verstehe, dass das Ziel dieser Studie eine Umfrage zur Lebensqualität und zu den allgemeinen Interessen der Teilnehmer ist UND KEINE WERBEMASSNAHME DARSTELLT.

- Ich stimme zu, dass alles, was ich während dieser Studie sehe oder lese, vertraulich behandelt werden sollte. Alle im Rahmen dieser Studie präsentierten Informationen dienen ausschließlich der Erforschung der Antworten auf die gestellten Fragen. Sie dürfen nicht dazu verwendet werden, Entscheidungen außerhalb des Forschungsumfelds zu beeinflussen.

- Ich verstehe, dass ich das Recht habe, die Beantwortung von Fragen zu verweigern oder jederzeit aus der Studie auszusteigen. Weitere Informationen zu meinen Rechten und zur Aufbewahrungsfrist meiner Daten finden Sie hier. [PN: CREATE HYPERLINK IN ‘here’ & PIPE: https://lifepointspanel.com/privacy for PATIENTS]

- Ich verstehe, dass alle von mir offengelegten Informationen streng vertraulich behandelt werden und die Ergebnisse der Forschung aggregiert werden, um ein Gesamtbild der Einstellungen zu den in dieser Umfrage behandelten Bereichen zu erhalten. Ich bleibe anonym, es sei denn, ich gebe meine Zustimmung zur Identifizierung. Meine persönlichen Daten werden ohne meine Zustimmung nicht an andere Organisationen weitergegeben.

Ich bestätige, dass ich die oben genannten Punkte gelesen und verstanden habe und ihnen zustimme und bin bereit, auf dieser Grundlage an der Forschungsumfrage teilzunehmen.

Klicken Sie auf die folgenden Links, um die vollständigen Datenschutzbestimmungen und Allgemeinen Geschäftsbedingungen anzuzeigen.

(1)	(2)
Annehmen	Ablehnen **[PN: TERMINATE IF SELECTED]**

[PN: INSERT BELOW HYPERLINKS INTO WORDS ABOVE, DO NOT SHOW THESE LINKS LIKE THIS]

DATENSCHUTZERKLÄRUNG FÜR PATIENTENSTUDIEN:
https://lifepointspanel.com/privacy

ALLGEMEINE GESCHÄFTSBEDINGUNGEN FÜR PATIENTENSTUDIEN:
https://www.lifepointspanel.com/terms-of-service

[SCREEN 5 – PN: NEW SCREEN. SHOW FOR ALL PATIENT STUDIES. SINGLE CODE – MAINTAIN THIS ORDER]

Diese Umfrage befasst sich mit Ihrer Gesundheit und Ihrem Gesundheitszustand. Alle in dieser Umfrage erhobenen sensiblen Daten werden gemäß unserer Datenschutzrichtlinie vertraulich behandelt. Dies ist ein sensibles Thema, das manchen Menschen Unbehagen bereiten könnte. Wenn Ihnen die Beantwortung von Fragen zu diesem Thema unangenehm ist, können Sie die Umfrage jederzeit jetzt oder während der Umfrage schließen.

Sind Sie bereit, freiwillig an dieser Studie teilzunehmen?

1. Ja, ich bin mit der Teilnahme einverstanden.

NEXT SCREEN

2. Nein, ich bin mit der Teilnahme nicht einverstanden.

S1. Welche der folgenden Aussagen trifft am besten auf Sie zu:

1. Bei mir wurde in den letzten 3 Monaten zum ersten Mal Krebs diagnostiziert. (CONTINUE)
2. Ich befinde mich derzeit in Krebstherapie. (CONTINUE)
3. Ich bin in Remission / Ich habe meinen Krebs besiegt. (CONTINUE)
4. Bei mir wurde noch nie Krebs diagnostiziert. (CLOSE)
5. Anderes (CLOSE)

INTRODUCTION ON AN EXTRA PAGE:

Wir sind an einigen Angaben interessiert, die Sie und Ihre Gesundheit betreffen. Bitte beantworten Sie die folgenden Fragen selbst, indem Sie die Zahl einkreisen, die am besten auf Sie zutrifft. Es gibt keine „richtigen“ oder „falschen“ Antworten. Ihre Angaben werden streng vertraulich behandelt.

THE FOLLOWING ANSWERS SHOULD BE RADIO BUTTONS WITH THE NUMBER NEXT TO IT:

		Überhaupt nicht	Wenig	Ziemlich	Sehr
C1.	Bereitet es Ihnen Schwierigkeiten, sich körperlich anzustrengen (z. B. eine schwere Einkaufstasche oder einen Koffer zu tragen)?	1	2	3	4
C2.	Bereitet es Ihnen Schwierigkeiten, einen längeren Spaziergang zu machen?	1	2	3	4
C3.	Bereitet es Ihnen Schwierigkeiten, eine kurze Strecke außer Haus zu gehen?	1	2	3	4
C4.	Müssen Sie tagsüber im Bett liegen oder in einem Sessel sitzen?	1	2	3	4
C5.	Brauchen Sie Hilfe beim Essen, Anziehen, Waschen oder Benutzen der Toilette?	1	2	3	4

Während der letzten Woche:	**Überhaupt nicht**	**Wenig**	**Ziemlich**	**Sehr**
C6. Waren Sie bei Ihrer Arbeit oder bei anderen tagtäglichen Beschäftigungen eingeschränkt?	1	2	3	4
C7. Waren Sie bei Ihren Hobbys oder anderen Freizeitbeschäftigungen eingeschränkt?	1	2	3	4
C8. Waren Sie kurzatmig?	1	2	3	4
C9. Hatten Sie Schmerzen?	1	2	3	4
C10. Mussten Sie sich ausruhen?	1	2	3	4
C11. Hatten Sie Schlafstörungen?	1	2	3	4
C12. Fühlten Sie sich schwach?	1	2	3	4
C13. Hatten Sie Appetitmangel?	1	2	3	4
C14. War Ihnen übel?	1	2	3	4
C15. Haben Sie erbrochen?	1	2	3	4
C16. Hatten Sie Verstopfung?	1	2	3	4
C17. Hatten Sie Durchfall?	1	2	3	4
C18. Waren Sie müde?	1	2	3	4
C19. Fühlten Sie sich durch Schmerzen in Ihrem alltäglichen Leben beeinträchtigt?	1	2	3	4
C20. Hatten Sie Schwierigkeiten, sich auf etwas zu konzentrieren, z. B. auf Zeitunglesen oder Fernsehen?	1	2	3	4
C21. Fühlten Sie sich angespannt?	1	2	3	4
C22. Haben Sie sich Sorgen gemacht?	1	2	3	4
C23. Waren Sie reizbar?	1	2	3	4
C24. Fühlten Sie sich niedergeschlagen?	1	2	3	4
C25. Hatten Sie Schwierigkeiten, sich an Dinge zu erinnern?	1	2	3	4
C26. Hat Ihr körperlicher Zustand oder Ihre medizinische Behandlung Ihr Familienleben beeinträchtigt?	1	2	3	4
C27. Hat Ihr körperlicher Zustand oder Ihre medizinische Behandlung Ihr Zusammensein oder Ihre gemeinsamen Unternehmungen mit anderen Menschen beeinträchtigt?	1	2	3	4
C28. Hat Ihr körperlicher Zustand oder Ihre medizinische Behandlung für Sie finanzielle Schwierigkeiten mit sich gebracht?	1	2	3	4

THE FOLLOWING 2 QUESTIONS SHOULD USE A SLIDING SCALE FOR THE ANSWER

Bitte wählen Sie bei den folgenden Fragen die Zahl zwischen 1 und 7 ein, die am besten auf Sie zutrifft:

C29. Wie würden Sie insgesamt Ihren Gesundheitszustand während der letzten Woche einschätzen?

1	2	3	4	5	6	7
sehr schlecht						ausgezeichnet

C30. Wie würden Sie insgesamt Ihre Lebensqualität während der letzten Woche einschätzen?

1	2	3	4	5	6	7
sehr schlecht						ausgezeichnet

	Überhaupt nicht	**Wenig**	**Ziemlich**	**Sehr**
E1. In welchem Ausmaß fühlten Sie sich durch Nebenwirkungen Ihrer Behandlung belästigt?	1	2	3	4

D1. Welcher Krebs wurde bei Ihnen diagnostiziert? SINGLE ANSWER, RADIO BUTTONS

1. Brust
2. Anderer gynäkologischer Tumor
3. Lunge
4. Prostata
5. Leukämie
6. Magen
7. Speiseröhre
8. Darm
9. Haut (Melanom)
10. Leber
11. Hirn
12. Anderer Krebs, bitte spezifizieren Sie: OPEN END TEXT

D2. Wann wurde bei Ihnen die Krebsdiagnose gestellt? SINGLE ANSWER, RADIO BUTTONS

1. Innerhalb der letzten 6 Monate

2. Innerhalb des letzten Jahres
3. Innerhalb der letzten 2 Jahre
4. Innerhalb der letzten 5 Jahre
5. Innerhalb der letzten 10 Jahre
6. Vor mehr als 10 Jahren

D3. Sind Sie aktuell unter Behandlung, d.h. nehmen Sie irgendeine Art von Medikation oder bekommen Sie irgendeine Art von Behandlung gegen den Krebs? (Mehrfachantworten sind möglich) MULTIPLE ANSWERS POSSIBLE

1. Ja, ich werde aktuell mit Chemotherapie behandelt
2. Ja, ich werde aktuell mit Radiotherapie behandelt
3. Ja, ich werde aktuell mit Immun- bzw. zielgerichteter Therapie behandelt
4. Ja, ich hatte eine Operation innerhalb der letzten 3 Monate
5. Andere Art von Therapie, bitte spezifizieren Sie: OPEN END TEXT
6. Ich befinde mich derzeit in Nachsorge (mit regulären oder gelegentlichen Arztbesuchen)
7. Nein, ich bin derzeit nicht in Behandlung SINGLE CODE
8. Die Frage betrifft mich nicht SINGLE CODE

D4. Haben Sie noch weitere Erkrankungen? SINGLE ANSWER, RADIO BUTTONS

1. Nein
2. Ja

Falls ja, welche: QUESTION AND ANSWERS ONLY VISIBLE IF “Ja” WAS SELECTED; MULTIPLE ANSWERS POSSIBLE

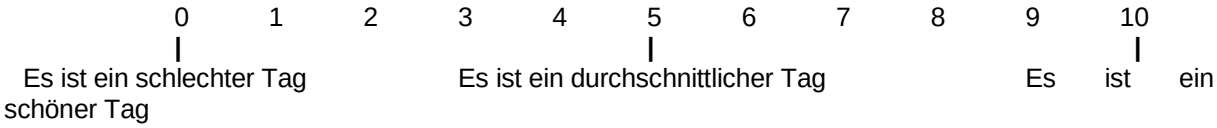
1. Nierenerkrankung
2. Herzerkrankung
3. Atemwegserkrankung
4. Rheumatische Erkrankung
5. Diabetes
6. Lebererkrankung
7. Andere Krankheit, bitte spezifizieren Sie: OPEN END TEXT

D5. Wie würden Sie Ihren aktuellen Aktivitätsgrad beschreiben? SINGLE ANSWER, RADIO BUTTONS

1. Voll aktiv und fähig, alle Tätigkeiten ohne Einschränkung durchzuführen
2. Aktiv, mit leichter Einschränkung bei körperlich anstrengenden Tätigkeiten
3. Einschränkungen in der Aktivität und bei körperlich anstrengenden Tätigkeiten
4. Selbstversorgung möglich, aber nicht fähig, einer Arbeitstätigkeit nachzugehen
5. Eingeschränkte Selbstversorgung mit mehr als 50% des Tages im Bett oder im Stuhl

ANSWER OF THE NEXT QUESTION SHOULD BE A SLIDING SCALE STARTING AT 0, ENDING AT 10. 0, 5 AND 10 SHOULD BE DENOTED WITH THE 3 WELL-BEINGS.

D6. Insgesamt betrachtet, wie geht es Ihnen heute? Bitte wählen Sie die Zahl, welche ihr aktuelles □efinden am besten beschreibt.



D7. Wie alt sind Sie? INTEGER FIELD (years)

D8. Welches Geschlecht haben Sie? SINGLE ANSWER, RADIO BUTTONS

1. Männlich
2. Weiblich
3. Weder noch
4. Keine Angabe

D9. Was ist ihr höchster Bildungsabschluss? SINGLE ANSWER, RADIO BUTTONS

1. Real- oder Mittelschule
2. Gymnasium
3. Universitätsabschluss
4. Promotion
5. Anderer, bitte spezifizieren Sie: OPEN END TEXT
6. Keine Angabe

D10. Wie ist ihr aktueller Beschäftigungsstatus? SINGLE ANSWER, RADIO BUTTONS

1. Vollzeit Student/in
2. Teilzeit Student/in
3. In Rente/Pension
4. Vollzeit Angestellte/r
5. Teilzeit Angestellte/r
6. Arbeitslos
7. Hausmann/Hausfrau
8. Geschäftsinhaber/in
9. Keine Angabe

D11. Wie ist Ihre aktuelle Lebenssituation? SINGLE ANSWER, RADIO BUTTONS

1. Ich lebe alleine
2. Ich lebe zusammen mit meinem Partner
3. Ich lebe zusammen mit meiner Familie (Ehepartner/Kinder/Eltern)
4. Ich lebe zusammen mit Freunden/anderen Personen
5. Ich lebe in einem Altersheim
6. Andere Lebenssituation, bitte spezifizieren Sie: OPEN END TEXT
7. Keine Angabe

D12. Treiben Sie aktiv Sport? SINGLE ANSWER, RADIO BUTTONS

1. Nein, gar nicht
2. Ja, gelegentlich
3. Ja regelmäßig, ca. einmal pro Monat
4. Ja regelmäßig, ca. einmal pro Woche
5. Ja, nahezu täglich

D13. Welchen Sport treiben Sie gerne? (Mehrfachantworten sind möglich) MULTIPLE ANSWERS POSSIBLE

1. Ski fahren
2. Wandern
3. Fahrrad fahren
4. Schwimmen
5. Bergsteigen
6. Gymnastik
7. Yoga
8. Reiten
9. Golfen
10. Anderer Sport, bitte spezifizieren Sie: OPEN END TEXT
11. Ich treibe keinen Sport SINGLE CODE

D14. Was ist Ihre Lieblingsfarbe? SINGLE ANSWER, RADIO BUTTONS

1. Rot
2. Blau
3. Grün
4. Gelb
5. Schwarz
6. Braun
7. Weiß
8. Andere Farbe, bitte spezifizieren Sie: OPEN END TEXT

D15. Welche Art von Freizeitbeschäftigung machen Sie am liebsten? (Mehrfachantworten sind möglich) MULTIPLE ANSWERS POSSIBLE

1. Lesen
2. Fernsehen
3. Im Internet surfen
4. Spazieren gehen
5. Sport
6. Zeit mit meiner Familie und Freunden verbringen
7. Kochen
8. Andere Aktivitäten, bitte spezifizieren Sie: OPEN END TEXT

D16. Welche Filme sehen Sie am liebsten? (Mehrfachantworten sind möglich) MULTIPLE ANSWERS POSSIBLE

1. Western
2. Krimis
3. Romantische Filme
4. Actionfilme
5. Animierte Filme
6. Komödien
7. Andere Filme, bitte spezifizieren Sie: OPEN END TEXT

D17. Welche Sportarten schauen Sie am liebsten am Fernseher? (Mehrfachantworten sind möglich) MULTIPLE ANSWERS POSSIBLE

1. Fußball
2. Eishockey
3. Skifahren
4. Leichtathletik
5. Fahrradfahren
6. Motorsport
7. Andere Sportarten, bitte spezifizieren Sie: OPEN END TEXT
8. Ich schaue kein Sport SINGLE CODE

D18. Wenn Sie einkaufen gehen, was kaufen Sie gerne ein? (Mehrfachantworten sind möglich) MULTIPLE ANSWERS POSSIBLE

1. Schuhe
2. Kleidung
3. Parfüm
4. Sportartikel
5. Küchenartikel
6. Lebensmittel
7. Haus- und Gartenartikel
8. Andere Dinge, bitte spezifizieren Sie: OPEN END TEXT
9. Ich gehe nicht gerne einkaufen SINGLE CODE

D19. Welche Art von Musik hören Sie häufig an? (Mehrfachantworten sind möglich) MULTIPLE ANSWERS POSSIBLE

1. Klassik
2. Rock
3. Folk
4. Jazz
5. Opern
6. Pop
7. Traditionell
8. Andere Musik, bitte spezifizieren Sie: OPEN END TEXT
9. Ich singe gerne oder musiziere lieber selbst, anstatt Musik zu hören SINGLE CODE
10. Musik interessiert mich nicht wirklich SINGLE CODE

D20. Was hatten Sie gestern zu essen?

- Mittagessen: OPEN END TEXT
- Abendessen: OPEN END TEXT

D21. Was ist Ihr Lieblingsessen? OPEN END TEXT

D22. Welche der folgenden Tiere haben Sie, sowohl im Zoo als auch in freier Wildbahn, in Ihrem Leben schon mal gesehen? (Mehrfachantworten sind möglich) MULTIPLE ANSWERS POSSIBLE

1. Elefant
2. Giraffe
3. Löwe
4. Leopard
5. Rhinozeros
6. Pinguin
7. Bär
8. Strauß
9. Adler
10. Andere Tiere, bitte spezifizieren Sie: OPEN END TEXT
11. Keine SINGLE CODE

D23. Lesen Sie aktuell ein Buch?

1. Nein
2. Ja, bitte nennen Sie den Titel des Buches, welches Sie gerade lesen: OPEN END TEXT

D24. Zu welcher Zeit gehen Sie üblicherweise ins Bett? OPEN END TEXT

D25. Welche Art von Urlaub bevorzugen Sie? (Mehrfachantworten sind möglich) MULTIPLE ANSWERS POSSIBLE

1. Strand: In der Sonne liegen und schwimmen
2. Berge: Wandern und spazieren gehen
3. Besichtigungstouren: Landschaften und Sehenswürdigkeiten
4. Aktiver Urlaub mit viel Sport
5. Zu Hause bleiben
6. Andere Art von Urlaub, bitte spezifizieren Sie: OPEN END TEXT

D26. Was war Ihr Lieblingsfach in der Schule? OPEN END TEXT

D27. Welcher ist der größte Planet in unserem Sonnensystem? SINGLE ANSWER, RADIO BUTTONS

1. Merkur
2. Venus
3. Erde
4. Mars
5. Jupiter
6. Saturn
7. Uranus
8. Neptun
9. Keine Ahnung

D28. Nennen Sie bitte 3 Länder, welche Sie interessant finden und in die Sie gerne reisen würden:

1. OPEN END TEXT
2. OPEN END TEXT
3. OPEN END TEXT

ANSWERS 2) AND 3) ARE NOT MANDATORY

D29. Welche Sprachen sprechen Sie neben Ihrer Muttersprache noch? OPEN END TEXT

D30. Wenn Sie eine Sprache wählen müssten, welche Sprache würden Sie gerne erlernen? OPEN END TEXT

D32. Was sind Ihre Lieblings-Automarken? (Mehrfachantworten sind möglich) MULTIPLE ANSWERS POSSIBLE

1. Mercedes
2. Jaguar

3. BMW
4. Volkswagen
5. Citroen
6. Chevrolet
7. Ford
8. Toyota
9. Opel
10. Landrover
11. Mini
12. FIAT
13. Andere Automarken, bitte spezifizieren Sie: OPEN END TEXT
14. Ich habe keine Lieblings-Automarke. SINGLE CODE

D32. Welche IT-Geräte nutzen Sie prinzipiell? (Mehrfachantworten sind möglich) MULTIPLE ANSWERS POSSIBLE

1. Smartphone
2. Tablet/iPad
3. PC/Notebook
4. Smartwatch
5. Andere IT-Geräte, bitte spezifizieren Sie: OPEN END TEXT

D33. Nehmen wir das Gerät, welches Sie am häufigsten benutzen. Wie regelmäßig nutzen Sie dieses Gerät? SINGLE ANSWER, RADIO BUTTONS

1. Täglich
2. Mehrmals pro Woche
3. Einmal pro Woche
4. Einmal pro Monat
5. Weniger als einmal pro Monat

THE FOLLOWING ANSWERS SHOULD BE RADIO BUTTONS WITH THE NUMBER NEXT TO IT

Wir kommen jetzt zu dem letzten Frageblock. Auch wenn Ihnen ein paar Fragen bekannt vorkommen, bitte beantworten Sie alle Fragen bis zum Ende.

	Überhaupt nicht	Wenig	Ziemlich	Sehr
F1. Bereitet es Ihnen Schwierigkeiten, sich körperlich anzustrengen (z. B. eine schwere Einkaufstasche oder einen Koffer zu tragen)?	1	2	3	4
F2. Bereitet es Ihnen Schwierigkeiten, einen längeren Spaziergang zu machen?	1	2	3	4
F3. Bereitet es Ihnen Schwierigkeiten, eine kurze Strecke außer Haus zu gehen?	1	2	3	4
F4. Müssen Sie tagsüber im Bett liegen oder in einem Sessel sitzen?	1	2	3	4
F5. Brauchen Sie Hilfe beim Essen, Anziehen, Waschen oder Benutzen der Toilette?	1	2	3	4

Während der letzten Woche:

	Überhaupt nicht	Wenig	Ziemlich	Sehr
F6. Waren Sie bei Ihrer Arbeit oder bei anderen tagtäglichen Beschäftigungen eingeschränkt?	1	2	3	4
F7. Waren Sie bei Ihren Hobbys oder anderen Freizeitbeschäftigungen eingeschränkt?	1	2	3	4
F8. Hatten Sie Schwierigkeiten, sich auf etwas zu konzentrieren, z. B. auf Zeitunglesen oder Fernsehen?	1	2	3	4
F9. Fühlten Sie sich angespannt?	1	2	3	4
F10. Haben Sie sich Sorgen gemacht?	1	2	3	4
F11. Waren Sie reizbar?	1	2	3	4
F12. Fühlten Sie sich niedergeschlagen?	1	2	3	4
F13. Hatten Sie Schwierigkeiten, sich an Dinge zu erinnern?	1	2	3	4
F14. Hat Ihr körperlicher Zustand oder Ihre medizinische Behandlung Ihr Familienleben beeinträchtigt?	1	2	3	4
F15. Hat Ihr körperlicher Zustand oder Ihre medizinische Behandlung Ihr Zusammensein oder Ihre gemeinsamen Unternehmungen mit anderen Menschen beeinträchtigt?	1	2	3	4

THE FOLLOWING 2 QUESTIONS SHOULD USE A SLIDING SCALE FOR THE ANSWER

Bitte wählen Sie bei den folgenden Fragen die Zahl zwischen 1 und 7 ein, die am besten auf Sie zutrifft:

F16. Wie würden Sie insgesamt Ihren Gesundheitszustand während der letzten Woche einschätzen?

1	2	3	4	5	6	7
sehr schlecht						ausgezeichnet

F17. Wie würden Sie insgesamt Ihre Lebensqualität während der letzten Woche einschätzen?

1	2	3	4	5	6	7
sehr schlecht						ausgezeichnet

NEW SCREEN

E2. Wir sind nun am Ende der Studie angekommen. Gerne können Sie uns hier noch weitere Kommentare/Anmerkungen mitteilen. OPEN END TEXT. NOT MANDATORY

TABELLENVERZEICHNIS

Tabelle 1: Rekrutierungsziele 29
Tabelle 2: Aufbau und Skalen des EORTC QLQ-C30 30
Tabelle 3: Übersicht zur Überprüfung und Bewertung der Modellannahmen einer linearen Regression 49
Tabelle 4: Kennzahlen zur Patientenrekrutierung 56
Tabelle 5: Qualitätsüberprüfung der Daten basierend auf gegensätzlichen Antworten bei den ersten 7 Fragen 57
Tabelle 6: Qualitätsüberprüfung der Daten basierend auf den Antwortzeiten 59
Tabelle 7: Definition der Analysepopulation basierend auf den Datenqualitäts-überprüfungen 61
Tabelle 8: Patienten Charakteristiken (n=2.643) 63
Tabelle 9: Dauer der Erhebung der einzelnen Fragebögen, der einzelnen Fragen und der gesamten Umfrage 66
Tabelle 10: Ergebnisse der DIF-Analyse 68
Tabelle 11: Art und Interpretation der detektierten DIF-Items 70
Tabelle 12: Übereinstimmung auf Item-Ebene und gewichtetes Kappa (Test-Retest-Reliabilität auf Item-Ebene) 80
Tabelle 13: Signifikante DIF-Items in Abhängigkeit der Symptomlast 82
Tabelle 14: Signifikante DIF-Items der Subgruppen Männer und Frauen 84
Tabelle 15: Standardisierte Faktorladungen der konfirmatorischen Faktorenanalyse des EORTC QLQ-F17 (N = 1.323, erste Erhebung) und des QLQ-C30 (N = 1.320, erste Erhebung) 85
Tabelle 16: Modell-Fit-Indizes der konfirmatorischen Faktorenanalyse des QLQ-F17 (N = 1.323, erste Erhebung) und des QLQ-C30 (N = 1.320, erste Erhebung) 86
Tabelle 17: Interfaktorkorrelationen der konfirmatorischen Faktorenanalyse des EORTC QLQ-F17 (N = 1.323, erste Erhebung) und des QLQ-C30 (N = 1.320, erste Erhebung) 87
Tabelle 18: Interne Konsistenz, konvergente Validität, diskriminante Validität und Skalierungsfehler des QLQ-F17 (N = 1.323, erste Erhebung) und des QLQ-C30 (N = 1.320, erste Erhebung) 88
Tabelle 19: Korrigierte Item-Skala-Korrelationen des QLQ-F17 88

Tabelle 20: EORTC QLQ-C30 Mittelwerte der Funktionsskalen im Vergleich zu Referenzwerten nach Altersgruppen und insgesamt 90
Tabelle 21: EORTC QLQ-C30 Mittelwerte der Symptomskalen im Vergleich zu Referenzwerten nach Krebsstatus und insgesamt 91

ABBILDUNGSVERZEICHNIS

Abbildung 1: Modularer Ansatz der EORTC zur Erhebung von Lebensqualität 16
Abbildung 2: Aufbau des QLQ-F17 17
Abbildung 3: Hypothesenformulierung bei Äquivalenzstudien 20
Abbildung 4: Übersicht Studienablauf 28
Abbildung 5: Fallzahlberechnung mit SAS 9.4 37
Abbildung 6: Beispiele zur Beurteilung von Äquivalenzstudien 47
Abbildung 7: Vergleich der Antworten der ersten 7 Items des QLQ-F17 und des QLQ-C30 58
Abbildung 8: Histogramm zur Gesamtbearbeitungszeit der Befragung. 59
Abbildung 9: Histogramm zur Bearbeitungszeit des QLQ-C30 unterteilt nach Zeitpunkt der Erhebung 60
Abbildung 10: Histogramm zur Bearbeitungszeit des QLQ-F17 unterteilt nach Zeitpunkt der Erhebung 60
Abbildung 11: Flussdiagramm des Patientenverlaufs 62
Abbildung 12: Diagnostische Diagramme für das DIF-Item C20 / F8 71
Abbildung 13: Diagnostische Diagramme für das DIF-Item C21 / F9 72
Abbildung 14: Diagnostische Diagramme für das DIF-Item C22 / F10 73
Abbildung 15: Diagnostische Diagramme für das DIF-Item C26 / F14 74
Abbildung 16: Diagnostische Diagramme für das DIF-Item C27 / F15 75
Abbildung 17: Forest Plot der multiplen linearen Regression im Zwischen-Gruppenvergleich 76
Abbildung 18: Forest Plot der linear gemischten Modelle im Innerhalb-Gruppenvergleich 78
Abbildung 19: Mittelwerte und 95%-Konfidenzintervalle der Skalen des QLQ-F17 und des QLQ-C30 in Abhängigkeit vom Zeitpunkt der Erhebung. 79
Abbildung 20: Intraklassen-Korrelationkoeffizienten (ICC) inkl. 95%-Konfidenz-intervall zwischen dem QLQ-C30 und dem QLQ-F17 für jede Skala 81

ABKÜRZUNGSVERZEICHNIS

Abkürzung	Bedeutung
CF	Cognitive Functioning (Kognitive Funktion)
CFI	Comparative Fit Index
DIF	Differential Item Functioning (Differentielle Item-Funktion)
EF	Emotional Functioning (Emotionale Funktion)
EMA	European Medicines Agency (Europäische Arzneimittel-Agentur)
EORTC	European Organisation for Research and Treatment of Cancer
FDA	Food and Drug Administration
GRM	Graded Response Mode
HRQOL	Health-Related Quality of Life (Gesundheitsbezogene Lebensqualität)
ICC	Intraclass Correlation Coefficient (Intraklassen Korrelationskoeffizient)
IRT	Item-Response-Theorie
ISO	International Organization for Standardization (Internationale Organisation für Normung)
KFA	Konfirmatorische Faktorenanalyse
KI	Konfidenzintervall
LMM	Linear Mixed Model (Linear gemischtes Model)
LR-Test	Likelihood-Ratio-Test
MCID	Minimal Clinically Important Difference (minimale klinisch relevante Differenz)
mm:ss	Minuten:Sekunden
PF	Physical Functioning (Körperliche Funktion)
PRO	Patient-Reported Outcome
QL	Global health status/Quality of Life (Globaler Gesundheitszustand/ Lebensqualität)
QLG	Quality of Life Group
QOL	Quality of Life (Lebensqualität)
RF	Role Functioning (Rollenfunktion)
RMSEA	Root Mean Square Error of Approximation

RS	Rohscore
SAP	Statistischer Analyseplan
SD	Standard Deviation (Standardabweichung)
SF	Social Functioning (Soziale Funktion)
SRMR	Standardized Root Mean Square Residual
TLI	Tucker-Lewis Index
UKR	Universitätsklinkum Regensburg
VIF	Variance Inflation Factor (Varianzinflationsfaktor)
WHO	World Health Organization
α	Alpha (Signifikanzniveau)
β	Beta (Fehler 2. Art)
θ	Theta (Latente Merkmalsausprägung)

DANKSAGUNG

An dieser Stelle möchte ich allen Menschen meinen großen Dank aussprechen, die mich bei der Anfertigung meiner Dissertation unterstützt haben.

Besonders danken möchte ich Prof. Michael Koller für das entgegengebrachte Vertrauen, mir die wissenschaftliche Leitung dieses Vorhabens zu übertragen, sowie für die kontinuierliche und hervorragende Betreuung und Unterstützung im gesamten Verlauf der Arbeit.

Ebenso danke ich Herrn Prof. Johannes Giesinger und Herrn Prof. Tobias Pukrop für ihre Betreuung als Mentoren und die fortwährende Unterstützung im Verlauf dieser Dissertation.

Mein größter Dank gilt meiner Frau Michaela, die mir mit unermüdlicher Geduld, Rückhalt und Verständnis den nötigen Freiraum zum Schreiben dieser Arbeit ermöglicht hat. Nicht zuletzt, indem sie sich in dieser Zeit liebevoll und selbstverständlich um unsere Kinder und das Alltagsleben gekümmert hat. Ohne diese Unterstützung wäre die Fertigstellung der Arbeit in dieser Form nicht möglich gewesen.

SELBSTSTÄNDIGKEITSERKLÄRUNG

“Ich, Zeman Florian geboren am 14.04.1982 in Weiden in der Oberpfalz, erkläre hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe.

Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Insbesondere habe ich nicht die entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten (Promotionsberater*in oder andere Personen) in Anspruch genommen.

Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.“

Ort, Datum Florian Zeman