Bähren, Tobias et al.

Research Report

Der Einsatz von grünem Tee und anderen Polyphenolen in der Medizin − eine Big-Data-Analyse der medizinischen Fachliteratur

ifid Schriftenreihe: Beiträge zu IT-Management & Digitalisierung, No. 1

Provided in Cooperation with:
ifid Institut für IT-Management & Digitalisierung, FOM Hochschule für Oekonomie & Management

Suggested Citation: Bähren, Tobias et al. (2021) : Der Einsatz von grünem Tee und anderen Polyphenolen in der Medizin − eine Big-Data-Analyse der medizinischen Fachliteratur, ifid Schriftenreihe: Beiträge zu IT-Management & Digitalisierung, No. 1, ISBN 978-3-89275-120-5, MA Akademie Verlags- und Druck-Gesellschaft mbH, Essen

This Version is available at:
https://hdl.handle.net/10419/249996

ifid Schriftenreihe
Beiträge zu IT-Management & Digitalisierung

Band 1

Rüdiger Buchkremer (Hrsg.)

Der Einsatz von grünem Tee und anderen Polyphenolen in der Medizin – eine Big-Data-Analyse der medizinischen Fachliteratur

~

Bähren / Braka / Burchard / Cyron / Demary / Dragieva / Eis / Farid / Gomes / Hacker / Kaiser / Krüger / Luu / Maasjosthusmann / Marks / Pachocki / Pongratz / Schade / Urban / Walter / Winter / Yesilyurt / Buchkremer

FOM Hochschule
ifid
Institut für IT-Management & Digitalisierung
der FOM University of Applied Sciences

Bähren / Braka / Burchard / Cyron / Demary / Dragieva / Eis / Farid / Gomes / Hacker / Kaiser / Krüger / Luu / Maasjosthusmann / Marks / Pachocki / Pongratz / Schade / Urban / Walter / Winter / Yesilyurt / Buchkremer

Der Einsatz von grünem Tee und anderen Polyphenolen in der Medizin – eine Big-Data-Analyse der medizinischen Fachliteratur

ifid Schriftenreihe der FOM, Band 1
Beiträge zu IT-Management & Digitalisierung

Essen 2021

ISBN (Print) 978-3-89275-119-9 ISSN (Print) 2699-562X
ISBN (eBook) 978-3-89275-120-5 ISSN (eBook) 2699-5638

Dieses Werk wird herausgegeben vom ifid Institut für IT-Management & Digitalisierung der FOM Hochschule für Oekonomie & Management gGmbH

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

MA Akademie Verlags-
und Druck-Gesellschaft mbH
Leimkugelstraße 6, 45141 Essen
info@mav-verlag.de

Rüdiger Buchkremer (Hrsg.)

Der Einsatz von grünem Tee und anderen Polyphenolen in der Medizin – eine Big-Data-Analyse der medizinischen Fachliteratur

Tobias Bähren / Daniel Braka / Patrick Burchard / Stanley Cyron / Marius Demary / Martina Dragieva / Luke Eis / Abdul Tanwir Farid / Daniel Gomes / Milan Hacker / Julian Kaiser / Robert Krüger / Simone Luu / Robin Maasjosthusmann / Alexander Marks / Christian Pachocki / Michael Pongratz / Johannes Christian Schade / Philipp Urban / Annika Walter / Viviane Winter / Erdal Yesilyurt und Rüdiger Buchkremer

Autorenkontakt:

E-Mail: ruediger.buchkremer@fom.de

Vorwort

Das ifid Institut für IT-Management & Digitalisierung der FOM Hochschule für Oekonomie & Management bündelt Forschungskompetenzen zu den Themen Künstliche Intelligenz, IT-Management und digitale Transformation. Dazu setzt das Institut auf die neuesten Erkenntnisse zu Big Data und KI-Analysemethoden. Die Aktivitäten der Forschenden des ifid werden durch Publikationen und Vorträge einer breiten Öffentlichkeit zugänglich gemacht, nun auch in der neugegründeten ifid Schriftenreihe.

Der erste Band mit dem Titel „Der Einsatz von grünem Tee und anderen Polyphenolen in der Medizin – eine Big-Data-Analyse der medizinischen Fachliteratur“ widmet sich der Analyse eines umfangreichen Korpus medizinischer Fachtexte. Im Rahmen ihrer Auswertung gehen die Autorinnen und Autoren der Frage nach, wie große Datenmengen durch Bündelung und Visualisierung der Informationen umfassend ausgewertet werden können. Impulsgebend für die weitere Forschung ist dabei nicht nur die Auswertung, sondern auch die Entstehung des Beitrags selbst. Die Autorinnen und Autoren fanden sich im Jahr 2019 im Rahmen meiner Lehrveranstaltung „Big-Data-Analyseprojekt“ des Masterstudiengangs „Big Data & Business Analytics“ an der FOM Hochschule in Düsseldorf zusammen. In verschiedenen Gruppen übernahmen die Studierenden u. a. Rollen in der Programmierung und der Visualisierung und trugen ihre Ergebnisse als Gemeinschaftsprojekt im vorliegenden Band zusammen.

Es ist das Ziel der neuen Schriftenreihe, herausragende Beiträge zu Themen der Digitalisierung und des IT-Management zu veröffentlichen. Dazu liefert der erste Band einen wertvollen Einstieg. Ich danke allen Beteiligten und hoffe auf eine breite Multiplikation in Lehre und Forschung.

Düsseldorf, im Oktober 2021

Prof. Dr. Rüdiger Buchkremer

Direktor des ifid Institut für IT-Management & Digitalisierung

Zusammenfassung

Um Anwendungsmöglichkeiten, Behandlungsgebiete und Wirkweisen der Inhaltsstoffe von grünem Tee zu erforschen, ist es erforderlich, zunächst einen Überblick über das existierende Fachwissen und bisherige Forschungsansätze zu erlangen. Dieses erweist sich jedoch als schwierig, wenn es zu einem Thema sehr viele Publikationen gibt, wie auch in diesem Beispiel zu grünem Tee. Unter Verwendung von Big-Data Analyse-Methoden und Künstlicher Intelligenz kann daher ein sogenannter „Korpus“ mit einer großen Anzahl relevanter Fachartikel erzeugt und analysiert werden. Die darauf aufbauende Analyse und Visualisierung vereinfacht die Exploration der zur Verfügung stehenden Informationen, wie die Erstellung einer Übersicht über die Hauptforschungsgebiete, die Erstellung einer Liste der bedeutenden Autorinnen und Autoren, die Auswertung der relevantesten Fachzeitschriften sowie die wichtigsten Anwendungsmöglichkeiten in Bezug auf Krankheiten und Symptome. In diesem Beitrag präsentiert das vorliegende Paper die Analyse und Visualisierung der Daten und Informationen von ca. 12.000 publizierten wissenschaftlichen Artikeln über grünen Tee aus der medizinischen Fachdatenbank PubMed.

Inhalt

Abkürzungsverzeichnis

API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
GTE	Camelia sinensis
EC	Epicatechin
EGC	Epigallocatechine
EGCG	Epigallocatechingallate
KI	Künstliche Intelligenz
MeSH	Medical Subject Headings
NCBI	National Center for Biotechnology Information
NLP	Natural Language Processing
ROS	Reaktive Sauerstoffspezies
SJR	Scimago Journal & Country
UI	Unique ID

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Wenn von der besonderen medizinischen Wirksamkeit des grünen Tees die Rede ist, ist häufig die Wirksamkeit eines bestimmten Wirkstoffes, des sogenannten Epigallocatechin Gallat, gemeint. Etwa 30 Prozent des Trockengewichts von Blättern des grünen Tees stellt die Substanz (-) Epigallocatechin-3-Gallat dar, welches auch mit EGCG abgekürzt wird (vgl. Brückner et al., 2012). Diese Substanz gehört zur Gruppe der Polyphenole, welche im Allgemeinen zu den Antioxidantien zählt (vgl. Liu, 2010).

Die potenziellen gesundheitlichen Vorteile des Konsums von grünem Tee und anderen Polyphenolen erlangen regelmäßig Aufmerksamkeit in der wissenschaftlichen Literatur. Die Anwendungsbereiche und Anwendungsmethoden variieren dabei sehr und reichen von der Behandlung von Krebs über Alzheimer bis zur Fettleibigkeit (Stevenson & Hurst, 2007). Im Allgemeinen zählt Tee weltweit zu den am häufigsten konsumierten Getränken nach Wasser und vor Kaffee. Vor allem in Ländern wie China, Japan, Teilen Nordafrikas und dem Nahen Osten hat insbesondere der Konsum von grünem Tee eine lange Tradition (vgl. Graham, 1992).

Generell besteht Tee aus den getrockneten Blättern der Teepflanze, Camellia sinensis (vgl. Yang & Landau, 2000). Es gibt unterschiedliche Herstellungsverfahren, welche zu den verschiedenen Teesorten führen. Nach Graham wird Grüntee im Gegensatz zu schwarzem Tee so hergestellt, dass die Oxidation der Polyphenole, welche den wichtigsten Bestandteil des Tees darstellen, bei der Trocknung der Blätter verhindert wird. Polyphenole sind sekundäre Pflanzenstoffe, zu welchen die Gruppe der Flavonoide zählt. Bestandteil dieser ist wiederum die Gruppe der Catechine, bei welchen es sich um Antioxidantien handelt, die im getrockneten Zustand bis zu 30 Prozent des Teeblatts ausmachen können. Die Catechine können in zwei Formen vorkommen, der größte Teil der Catechine des Teeblatts entspricht allerdings der Form Epicatechin, EC. Bei einer bestimmten molekularen Anordnung werden diese auch als Epigallocatechine, EGC, bezeichnet. Wenn diese zusätzlich mit der Verbindung Gallussäure Ester bilden, entstehen Epigallocatechingallate, EGCG (vgl. Graham, 1992). Eben diese Verbindungen sind in hoher Substanz in frischen Teeblättern enthalten und stellen gleichzeitig den wichtigsten Teil dar (vgl. Graham, 1992). In einer frisch zubereiteten Tasse grünem Tee sind bis zu 200 mg EGCG enthalten. Aufgrund der antioxidativen Eigenschaft der enthaltenen Catechine wurde die Anwendung von grünem Tee am häufigsten bei Krankheiten mit Verbindung zu sogenannten

freien oder Sauerstoffradikalen untersucht. Hinzu kommen weitere Anwendungsgebiete, bei welchen EGCG eine Rolle spielt, wie beispielsweise das Tumorwachstum, Parkinson, Alzheimer, Diabetes, Hypercholesterinämie sowie unspezifische antibakterielle und entzündungshemmende Einsatzgebiete (vgl. Zaveri, 2006). Zum Beispiel wurde eine Menge von 400 mg EGCG auf die medizinische Wirksamkeit bei traumatischen Hirnverletzungen durch klinische Studien getestet (vgl. Mao et al., 2018). Nach Yang und Landau werden in Tierstudien nicht selten höhere Dosierungen verabreicht, um die gesundheitlichen Vorteile prägnanter zu evaluieren. Bei der Einnahme großer Mengen an Grüntee beim Menschen ist unklar, inwiefern die enthaltenen Polyphenole aufgrund der hohen Bindungsaktivität zu ernährungsbedingten Problemen führen können (vgl. Yang & Landau, 2000). Die regelmäßige Einnahme in einer Größenordnung von 300 mg EGCG pro Tag wird als unbedenklich betrachtet; bei der Verabreichung von mehr als 800 mg pro Tag muss mit unangenehmen Nebenwirkungen gerechnet werden (vgl. Dekant et al., 2017).

Diese Arbeit verfolgt das Ziel, einen Überblick über die Vielzahl der existierenden Forschungsbereiche und -ergebnisse mit Bezug auf medizinische Anwendungsmöglichkeiten von grünem Tee und dessen Bestandteile sowie Erkenntnisse über den Verlauf von Publikationsdaten der relevanten Fachartikel mittels Big-Data-Analysen zu gewinnen.

Im Rahmen der Veranstaltung „Big-Data-Analyseprojekt“ des Masterstudiengangs „Big Data & Business Analytics“ an der FOM Hochschule in Düsseldorf im Jahr 2019 wurde diese Arbeit als gemeinsames Projekt erstellt. Die Teilnehmerinnen und Teilnehmer wurden in verschiedene Gruppen unterteilt, welche das Projektmanagement, die Programmierungs-, die Visualisierungs- sowie die Schreibgruppe umfassten.

2 Forschungsfragen

Im weiteren Verlauf dieser Arbeit sollen folgende Forschungsfragen geprüft werden:

- Ist es möglich, durch die Anwendung von Big-Data-Techniken auf wissenschaftliche Artikel einen umfassenden Überblick über Anwendungsbereiche und Krankheitsbehandlungen zu gewinnen?
- Welche Autoren beschäftigen sich mit der Anwendung von grünem Tee am häufigsten?
- Welche wissenschaftlichen Zeitschriften publizieren die meisten Artikel zum Einsatz von grünem Tee?
- Wie ist der zeitliche Verlauf der publizierten Fachartikel zum Thema „Therapie mit grünem Tee“?
- Bei welchen Krankheiten und Symptomen wird der Wirkstoff von grünem Tee erfolgreich angewendet?
- In welchen Bereichen der Medizin wird das Thema grüner Tee am häufigsten behandelt?

3 Theoretische Grundlagen

Zur Linderung bei Entzündungen gilt Grüntee seit 5.000 Jahren als beliebtes Getränk. In der Stammpflanze Camellia sinensis, GTE, sind unter anderem die Polyphenole EC, EGC und EGCG auffindbar. Den einflussreichsten Anteil und gleichzeitig knapp ein Drittel der gesamten Catechine weist das Polyphenol EGCG auf. Es ist erwiesen, dass EGCG entzündungshemmend gegenüber unterschiedlichen Formen von pro-inflammatorischen Faktoren wirkt (vgl. Hagiu et al., 2020).

Eine weitere Anwendung findet Grüntee oder EGCG bei Krebs (vgl. Surh, 2003). EGCG reagiert über verschiedene Wege mit sogenannten freien Radikalen (ROS). Es kann sowohl antioxidative als auch pro-oxidative Wirkungen entfalten. Die antioxidative Eigenschaft wirkt dabei wie folgt: Sie reduziert die ROS, was beispielsweise zur Apoptose von Krebszellen führen kann (vgl. Hayakawa et al., 2019). Studien belegen, dass eine umgekehrte Assoziation zwischen dem Konsum von grünem Tee und dem Krebsrisiko existiert. In 51 von 127 Fall-Kontroll-Studien und 19 von 90 Gruppenstudien wurde gezeigt, dass unter anderem folgende Krebsarten betroffen sind: Blasen-, Brust-, Darm-, Magen-, Nieren-, Lungen-, Eierstock-, Bauchspeicheldrüsen- und Prostatakrebs. Eine europäische Studie belegte, dass ein erhöhter Teekonsum die Entstehung eines Leberzellkarzinoms um 59 Prozent verringert. Diese Studie untersuchte 11 Jahre lang 486.799 Probanden, unabhängig von deren Geschlecht. Eine weitere Analyse zeigte, dass erhöhter Teekonsum mit einem reduzierten Risikoverhältnis von *0,72* auf das Auftreten von Mundkrebs notiert ist. Diese Analyse nahm als Grundlage 87 Datensätze aus 57 Studien, die 49.812 Probanden beinhaltete (vgl. Hayakawa et al., 2019).

Polyphenole sollen auch hilfreich bei Knochenschwund sein, wie Tierversuche gezeigt haben. Shen et al. prüften die Wirksamkeit von Substanzen in Kombination mit Alphacalcidol, um den Knochenverlust zu reduzieren. Als Probanden wurden Ratten mit chronischen Entzündungen verwendet. Die Ergebnisse zeigen, dass sich die Knochenmasse durch die Einnahme der Substanzen erhöht hatte (vgl. Shen et al., 2010).

Zum Thema Übergewicht wurden 48 erwachsene Mäuse unter eine Kontrolldiät gestellt. Eine der Kontrollgruppen erhielt zur Sucrose zusätzlich zwei Prozent Grünteepulver. Die Grünteegruppe weist folgende Eigenschaften im Gegensatz

zur Vergleichsgruppe auf: geringeres Körpergewicht, Körperfett, sowie geringerer Leberfettstatus und höhere magere Masse der Leber (vgl. Cichello et al., 2013).

Bluthochdruck ist bei erwachsenen Menschen stark verbreitet. Bis zum Jahr 2025 wird mit einer globalen Verbreitung von etwa 30 Prozent gerechnet. Tee trägt zu einer deutlichen Reduktion von Herz-Kreislauf-Erkrankungen bei. Das Risiko von Bluthochdruck wurde durch den Konsum von grünem Tee jedoch nicht reduziert (vgl. Chei et al., 2018).

4 Methodik und Untersuchungsgegenstand

Damit Informationen zu dem Themenkomplex mit Big-Data-Methoden analysiert werden können, muss eine maschinenlesbare Datenbasis als Grundlage geschaffen werden. Hierzu eignet sich eine Online-Recherche über PubMed, ein Portal, welches in erster Linie auf die MEDLINE-Datenbank mit Referenzen und Abstracts zu lebenswissenschaftlichen und biomedizinischen Themen zugreift. PubMed umfasst mehr als 30 Millionen Zitate über biowissenschaftliche Zeitschriften und zu Online-Büchern (vgl. US National Library of Medicine, 1996).

Zum Durchführen einer Datenanalyse ist die Erstellung eines Datenkorpus essenziell (vgl. Buchkremer et al., 2019), da dieser als Grundlage für weiterreichende Analysen dient. In Kapitel 4.1 wird zunächst die Erstellung des Daten-Korpus dargelegt, Kapitel 4.2 beschreibt die Bereitstellung von weiteren Daten anhand des vorliegenden Daten-Korpus (wie bspw. den Scimago Journal & Country Rank, SJR), schließlich erfolgt eine Häufigkeitsanalyse von Krankheiten und Symptomen in Kapitel 4.3, welche in Zusammenhand mit Grüntee erwähnt werden. Kapitel 4.4 überprüft des Weiteren die Entwicklung der Veröffentlichungen zu Grüntee von 1947 bis 2019, folgend wird in Kapitel 4.5 eine Analyse mittels BERT durchgeführt.

4.1 Erstellung des Daten-Korpus

Die Abfragen an die Metadatenbank PubMed können sowohl über den Webbrowser als auch per Programmierschnittstelle, API, erfolgen. Als Einstiegspunkt per Webbrowser dient die Metasuchmaschine Entrez, die eine breitgefächerte Suche erlaubt. Dabei wird der Einsatz von erweiterten Suchoperatoren ermöglicht und es werden vielfältige Werkzeuge zur Datenanalyse bereitgestellt (vgl. Buchkremer et al., 2019).

Die Artikel in PubMed werden zur Beschreibung des Themas mit Hilfe von biomedizinischen Fachbegriffen (Medical Subject Headings, MeSH) kategorisiert. MeSH-Begriffe sind dabei hierarchisch nach Fachkategorien geordnet, mit spezifischen Begriffen unterhalb von weitergefassten Begriffen. PubMed ermöglicht die Betrachtung der Hierarchie sowie die Verwendung von Begriffen zur MeSH- und Artikelsuche (vgl. Jimeno-Yepes et al., 2012).

Für die Analyse der gesundheitsfördernden Wirkungen von grünem Tee, unterstützt durch den Einsatz von Big Data, wird ein Datenkorpus benötigt (vgl. Buchkremer et al., 2019). Daher gilt es im ersten Schritt, eine Abfrage (engl.

query) zu erstellen, mit der eine möglichst große Anzahl an Treffern bei der PubMed-Suche zum Inhaltsstoff von Grüntee und dessen Synonymen gefunden wird. Hierzu muss eine größere Anzahl an Referenz-Informationen aus medizinischen Artikeln extrahiert werden, um einen Datenkorpus abzubilden, der in einer maschinenlesbaren Form für die Weiterverarbeitung benötigt wird. Die Referenz-Informationen müssen darüber hinaus um MeSH-Tree-Nummern, die ein Wortnetz darstellen, angereichert werden, damit eine Inhaltsklassifizierung und Sacherschließung in der weiteren Analyse ermöglicht wird.

Eine Suche auf PubMed mit dem Begriff „green tea“ ergibt bereits 8.203 Treffer. Um die Anzahl an Treffern zu steigern, werden weitere Begriffe mit dem initialen Suchbegriff über einen Oder-Operator verbunden. Bei den weiteren Suchbegriffen, wie epigallocatechin-3-gallate, epigallocatechin gallate und EGCG, handelt es sich um einen Bestandteil des Grüntees und dessen Synonyme aus der Gruppe der Catechine, denen eine gesundheitsfördernde Wirkung zugeschrieben wird (vgl. Ehrnhoefer et al., 2006).

Innerhalb von PubMed werden einzelne der bisher gefundenen Artikel über die XML-Ansicht betrachtet, über die sich chemische Substanzcodes, welche im Artikel verwendet werden, entnehmen lassen. So können zusätzlich zu den einfachen Suchbegriffen auch eindeutige Identifikationsnummern für Substanzen zur Suche hinzugefügt werden, um eine größere Trefferzahl zu erhalten.

Tabelle 1 zeigt die Anzahl der Treffer für unterschiedliche Queries mit den Sortierungen Best Match für eine größtmögliche Übereinstimmung und Most Recent für eine zeitliche Sortierung. Die Trefferanzahl für die ersten sechs Queries ist bei Most Recent mit geringen Unterschieden höher als bei Best Match. Allerdings zeigt sich bei den Queries 7 und 8, dass Best Match mehr Treffer liefert. In Query 7 liegt dies daran, dass sich innerhalb der Query ein Fehler befindet, der erst im späteren Verlauf bemerkt wird. Dabei ist der Suchbegriff „epigallocat*“ enthalten, in dem sich keine Anführungszeichen in Verbindung mit dem Stern-Operator befinden dürfen. Allerdings liefern sowohl Query 7 als auch 8 die identische und auch höchste Anzahl an Treffern mit der Sortierung nach Best Match, wobei Query 8 unter Verwendung von Most Recent die meisten Treffer aufweist.

Tabelle 1: Anzahl an Treffern mit unterschiedlichen Queries und Sortierung

Nr.	Query	Most Recent	Best Match
1	"green tea"	**8.203**	8.181
2	"green tea" OR "epigallocatechin-3-gallate"	**9.396**	9.375
3	"green tea" OR "epigallocatechin-3-gallate" OR "Epigallocatechin gallate"	**11.522**	11.497
4	"green tea" OR "epigallocatechin-3-gallate" OR "Epigallocatechin gallate" OR "EGCG"	**11.605**	11.580
5	"green tea" OR "epigallocatechin-3-gallate" OR "Epigallocatechin gallate" OR "EGCG" OR ("green tea extract AR25" [Supplementary Concept])	**11.605**	11.588
6	"green tea" OR "epigallocatechin-3-gallate" OR "Epigallocatechin gallate" OR "EGCG" OR ("green tea extract AR25" [Supplementary Concept]) OR ("epigallocatechin gallate" [Supplementary Concept])	**11.605**	11.588
7	"green tea" OR "epigallocatechin-3-gallate" OR "Epigallocatechin gallate" OR "EGCG" OR "epigallocat*"	11.605	**12.213**
8	"green tea" OR "epigallocatechin-3-gallate" OR "Epigallocatechin gallate" OR "EGCG" OR epigallocat*	12.210	**12.213**

Eine explorative Betrachtung der Benutzerführung von PubMed zeigt, dass sich Inhalte zwar per API abfragen lassen, jedoch nicht mit Sicherheit beantwortet werden kann, ob die benötigten Informationen wie MeSH-Tree-Nummern dabei

enthalten sind. Auch wenn sich ein automatisierter Zugriff über einen Webbrowser zur Datengewinnung als aufwendiger gestaltet, sind hierbei jedoch die benötigten Informationen enthalten.

Mithilfe von „Selenium“ (vgl. Muthukadan, 2019) kann ein Webbrowser automatisiert gesteuert werden, um so die Informationen zu sammeln und maschinenlesbar verfügbar zu machen. Eine dafür entwickelte Python-Applikation übernimmt dabei die Steuerung und enthält die Logik zur Datenextraktion. Die Problemformalisierung zur Extraktion der Artikel lautet wie folgt:

Es wird ein Algorithmus gesucht, mit dem sich die Einträge der Suche automatisiert aus PubMed extrahieren lassen.

Der Algorithmus wird wie in Abbildung 1 als Pseudocode definiert und innerhalb der Python-Applikation implementiert.

Abbildung 1: Algorithmus zur Datenextraktion der PubMed XML-Treffer

```
function load_articles:
        load pubmed-search
        switch to 'Most Recent'
        switch to '200 Items per Page'
        while(true):
                switch to 'XML-view'
                extract all text as XML-Tree
                for every 'PubmedArticle'
                        create file with content of the article
                click browser-back button --> back to summary-view
                if next-page-button does not exist
                        finish
                click next-page button --> shows the next 200 results
```

Die Problemformalisierung zur Anreicherung der Artikel mit MeSH-Tree-Nummern gestaltet sich wie folgt:

Es wird eine Methode gesucht, um die extrahierten XML-Dateien möglichst effizient um MeSH-Tree-Informationen anzureichern.

Die MeSH-Tree-Informationen befinden sich als eindeutige Identifikationsnummern (MeSH Unique ID) innerhalb der XML-Dateien. Dabei handelt es sich jedoch nicht um Identifikationsnummern, in denen Informationen über die medizinischen Fachgebietsüberschriften (MeSH) enthalten sind. Mithilfe der MeSH-ID

können allerdings die Webseiten, auf denen sich die MeSH-Tree-Nummern befinden, aufgerufen werden. Das Anzeigen der MeSH-Webseite erfolgt über die Suche von PubMed und der Such-Einstellung auf MeSH. Innerhalb der Seite sind die MeSH-Tree-Nummern als Tree-Number(s) gekennzeichnet und können hieraus extrahiert werden, um damit die XML-Datei anzureichern.

In Abbildung 2 ist der Algorithmus als Pseudocode dargestellt. Die Extraktion erfolgt dabei über eine automatisierte Steuerung mittels Selenium und stellt einen Teilprozess innerhalb der Python-Applikation dar.

Abbildung 2: Algorithmus zur Anreicherung der XML-Dateien um MeSH-Codes

```
function process():
        for every downloaded file:
                load_mesh_data(file)

function load_mesh_data(file):
        load file as xml
        for every 'MeshHeading':
                load mesh-tree-numbers from local-database
                if mesh-tree-numbers are not inside local-database
                        load mesh-tree-numbers from web
                        save mesh-tree-numbers in local-database
                add mesh-tree-numbers to corresponding mesh-code
```

Da die MeSH-IDs für die Zuordnung zu MeSH-Tree-Nummern immer eindeutig zu den gleichen Mesh-Tree-Nummern führen, kann jeder Artikel, der eine MeSH-ID enthält, auch immer mit den gleichen MeSH-Codes angereichert werden.

Innerhalb des Korpus befinden sich Dateien, welche die Informationen der PubMed-Referenz-Informationen und zusätzlichen MeSH-Tree-Nummern als XML-Dateien enthalten. Für eine Weiterverarbeitung wird ein Python-Reader der Artikel als separate Applikation erstellt. Dadurch lassen sich zum Beispiel die Titel oder das Veröffentlichungsdatum aller Abstracts ausgeben. Die Ausgabe einzelner Elemente eines Artikels über den gesamten Korpus hinweg ist insofern sinnvoll, als dass sich einzelne Bestandteile, wie die Abstracts der Artikel oder Veröffentlichungsdaten, genauer analysieren lassen.

Die Anzahl der durch die Suchabfragen gefundenen Referenz-Informationen umfasst 11.605 Treffer. Zwar enthält unser „Korpus“ nicht alle möglichen 12.210

Treffer, jedoch stellt die gefundene Anzahl eine Menge von 95,05 Prozent aller Treffer dar. Die nicht berücksichtigten Referenzen werden aufgrund der Feststellung von Fehlern (fehlende Daten, Dateninkonsistenz) nach der Analyse und Evaluation des gesamten Korpus nicht nachträglich hinzugefügt.

Die Extraktion der Referenz-Informationen als XML-Dateien erfolgt durch die Zuhilfenahme computergesteuerter Webbrowser automatisiert mittels einer Python-Applikation und ermöglicht somit eine effiziente und maschinenlesbare Informationsbereitstellung ohne manuellen Rechercheaufwand. Mit der Applikation wird weiterhin die Möglichkeit geschaffen, zukünftige Referenz-Informationen zu anderen Themen aus PubMed zu extrahieren, da die Implementierung unabhängig von den verwendeten Abfragen ist. Eine Anreicherung der Referenz-Informationen mit MeSH-Tree-Nummern kann mittels lokaler Datenbank als Cache effizient umgesetzt werden. Dadurch wird eine Anreicherung der Green-Tea-Daten durch MeSH-Tree-Nummern innerhalb eines Tages ermöglicht.

Die MeSH-Tree-Nummern stellen einen Thesaurus (altgriechisch: θησαυρός thesaurós, Schatz, Schatzhaus, lateinisch daher Thesaurus), bzw. ein Wortnetz dar, dessen Begriffe mittels Synonymen miteinander verbunden sind (vgl. Aitchison & Clarke, 2004). Mit ihrer Hilfe lassen sich in einer späteren Analyse Zusammenhänge zwischen Grüntee, weiteren Inhaltsstoffen und Krankheiten abbilden.

Die Erstellung eines Suchalgorithmus (Query), die Extraktion von Referenz-Informationen und die Anreicherung der Referenz-Informationen mit MeSH-Tree-Nummern mithilfe einer Python-Applikation führen zu einer Anzahl an 11.605 Dokumenten. Basierend auf diesem Korpus folgen weitere Analysen.

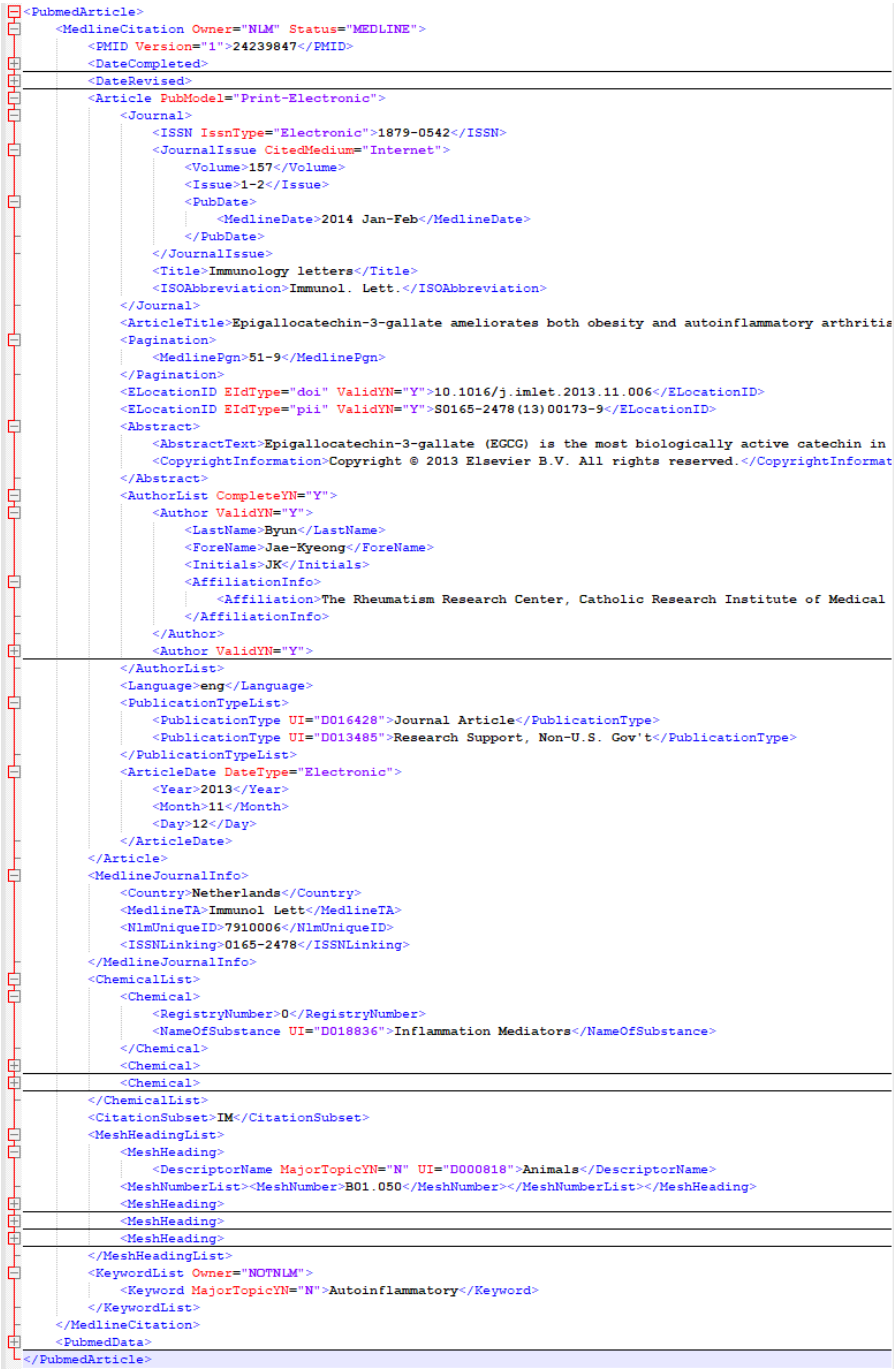
4.2 Bereitstellung von weiteren Daten

Als Grundlage für weitere Visualisierungen wurde festgestellt, dass zusätzliche Datenformate benötigt werden. Dies erfordert die Bereitstellung der für die Auswertungen relevanten Daten im CSV-Format. Zudem sollen die Daten aggregiert, gruppiert und wenn möglich um weitere Metadaten ergänzt werden. Außerdem soll der Einsatz von Big-Data-Analytics-Software evaluiert werden, damit nach Möglichkeit mittels künstlicher Intelligenz, KI, die Daten automatisiert analysiert werden können. Für die Entwicklung der Codes in diesem Kapitel wird Python in der Version 3.6 inklusive folgender Module verwendet:

- Datenhandling: Pandas
- XML: ElementTree
- SQL Lite: sqlite3
- Sonstige: os, datetime, logging, copyfile.

Der vorliegende Korpus enthält XML-Dateien, deren Aufbau in Abbildung 3 erkennbar ist.

Abbildung 3: Aufbau einer XML-Datei im Korpus



Alle Auswertungen werden nach manueller, stichprobenartiger Prüfung automatisiert über Python-Codes durchgeführt. Die Ergebnisse werden sowohl in einer SQLite-Datei als auch in Form von CSV-Dateien zur weiteren Verarbeitung bereitgestellt.

Als Hilfsmittel zur Ermittlung der Bezeichnungen (Eigenschaft: Name) einzelner MeSH-Codes wird eine Referenzliste verwendet. Unique IDs, UI, sind eindeutige Kennzeichen für einen Tag, welche einen Artikel des Korpus einer Kategorie zuweisen. Jeder Artikel kann mehrere Zuweisungen erhalten. Die neuesten Artikel können diesbezüglich nicht berücksichtigt werden, da die Zuweisung der Tags in der Datenbank PubMed manuell geschieht und dieser Prozess in einigen Fällen innerhalb des Projektzeitraumes noch nicht abgeschlossen war.

Zuweisungen von UIs im XML-Dokument sind im Attribut der Tags Descriptor-Name zu finden. Die darin enthaltene Zahlen- und Buchstabenkombination entspricht der UI. Abbildung 4 zeigt die Datenstruktur, in welcher die UI-Daten zusammengefasst werden.

Abbildung 4: Datenstruktur der Tabelle für Unique IDs

UIPerYear		CREATE TABLE UIPerYear (ui text, year int, name text, count int)
ui	text	"ui" text
year	int	"year" int
name	text	"name" text
count	int	"count" int

Alle Dokumente werden über ein Skript automatisiert durchlaufen und bezüglich der Angabe zur UI strukturiert. Über die Angabe zum Publikationsdatum wird das Jahr der Veröffentlichung bestimmt. Jedes Auftreten einer UI wird pro Jahr gezählt. Anschließend werden die Namen mittels der Referenzliste ergänzt.

Mit der Hilfe von UIs kann über einen Service von PubMed eingesehen werden, welche MeSH-Codes zu einer UI zusammengefasst sind. Dieser wurde während der Korpus-Erstellung bereits durchgeführt und demnach sind bereits alle MeSH-Codes pro UI im Dokument vorhanden. Zu finden sind diese in den Einträgen unter dem XML-Tag MeshNumberList.

Nachfolgende Abbildung zeigt die Datenstruktur der MeSH-Codes und MeSH-Code-Referenzen-Datei.

Abbildung 5: Datenstruktur der MeSH-Code-Tabellen

MeshCodes		CREATE TABLE MeshCodes (code text, name text, count int)
code	text	"code" text
name	text	"name" text
count	int	"count" int
MeshCodesPerYear		CREATE TABLE MeshCodesPerYear (code text, year int, name text, count int)
code	text	"code" text
year	int	"year" int
name	text	"name" text
count	int	"count" int
MeshCodesReferences		CREATE TABLE MeshCodesReferences (code1 text, code2 text)
code1	text	"code1" text
code2	text	"code2" text
MeshCodesReferencesCounts		CREATE TABLE MeshCodesReferencesCounts (code1 text, code2 text, count int)
code1	text	"code1" text
code2	text	"code2" text
count	int	"count" int

Für die MeSH-Codes selbst werden der Name und die Häufigkeit (einmal pro Jahr und einmal absolut) ermittelt. MeshCodesReferences zeigen die aufgetretenen Paarungen von MeSH-Codes. Pro Datei wird jeder Code zu jedem anderen Code als Paarung gespeichert. Über eine Zählung und Aufsummierung soll bestimmt werden, welche Paarungen besonders häufig auftreten. Daher wird schon beim Import gezählt, wie oft die Paarungen insgesamt vorkommen (MeshCodeReferencesCounts).

Abbildung 6 zeigt die Datenstruktur zum Ablegen der Index-Informationen zu den einzelnen Artikeln.

Abbildung 6: Datenstruktur der vollständigen MeSH-Code-Datei inkl. Journal und H-Index

FullMeshCodes		CREATE TABLE FullMeshCodes (pmid int, issn text, j
pmid	int	"pmid" int
issn	text	"issn" text
journal	text	"journal" text
countryBySjr	text	"countryBySjr" text
country	text	"country" text
sjrRank	int	"sjrRank" int
hIndex	float	"hIndex" float
year	int	"year" int
code	text	"code" text
journalBySjr	text	"journalBySjr" text

In der entsprechenden Datei finden sich die Daten aller Dokumente des Korpus. Mittels der pmid-Angabe, welche für die PubMed-ID steht, können die Artikel eindeutig identifiziert werden.

Ergänzt werden die Korpus-Daten durch Informationen des SCImago Journal & Country Rank, SJR, welcher unter anderem den H-Index pro Jahr und Journal bereitstellt (vgl. SCImago, 2019). Über die ISSN-Nummern können die Journals mit den externen Daten verknüpft und um die H-Index-Angaben ergänzt werden. Des Weiteren werden die Länderangabe (zum Abgleich) und der Rang des SJR als weiteres Metadatum aufgenommen.

Zusätzlich werden weitere Metadaten aufbereitet: die jeweilige Anzahl der Vorkommnisse von Autorinnen und Autoren sowie Journals pro Jahr und Land. Autoren werden mit Vor- und Nachnamen angegeben, Initialen werden nicht eingelesen. Zudem gibt es keine eindeutigen Identifizierungsmöglichkeiten für die Autoren, sodass im Sinne der Auswertung nicht immer sichergestellt werden kann, ob gleichnamige Autoren auch getrennt betrachtet werden. Abbildung 7 zeigt die Datenstruktur der Metadaten:

Abbildung 7: Datenstruktur der Metadaten-Tabellen

AuthorPerYear		CREATE TABLE AuthorPerYear (author text, year int, count int)
author	text	"author" text
year	int	"year" int
count	int	"count" int
JounalsAndAuthorsPerYear		CREATE TABLE JounalsAndAuthorsPerYear (pmid int, journal text, author text, country text, year int)
pmid	int	"pmid" int
journal	text	"journal" text
author	text	"author" text
country	text	"country" text
year	int	"year" int
JournalPerYear		CREATE TABLE JournalPerYear (journal text, year int, count int)
journal	text	"journal" text
year	int	"year" int
count	int	"count" int

Es wird hierbei einmal nach Autorinnen und Autoren und einmal nach Journal separat unterschieden und gezählt.

4.3 Häufigkeitsanalyse von Krankheiten und Symptomen

Um die Texte der Abstracts nach Krankheiten und Symptomen durchsuchen zu können, muss zunächst eine Liste von bekannten Begriffen erstellt werden. Nach Recherche wurde dabei die Arbeit von Zhou et al. als Grundlage gewählt (vgl. Zhou et al., 2014). In diesem Artikel erstellen die Autoren eine Übersicht aller in PubMed gelisteten Krankheiten und Symptome. Da geschultes Personal die Artikel von PubMed mit den entsprechenden MeSH-Codes markiert, ist es Zhou et al. möglich gewesen, in dieser Übersicht zu ermitteln, wie oft welcher MeSH-Code in der PubMed-Datenbank genannt wird. Hierbei wurde eine Übersicht für die Krankheiten und eine für die Symptome erstellt (vgl. Zhou et al., 2014).

Aus diesen Übersichtslisten werden die Namen der Krankheiten beziehungsweise der Symptome aus allen MeSH-Codes extrahiert, so dass jeweils eine Liste mit Krankheiten respektive Symptomen vorliegt. Zunächst wird ein Skript geschrieben, das die absolute Häufigkeit aller Krankheiten respektive Symptomen pro Abstract im Korpus auflistet. Hierbei durchläuft das Skript jeden Abstract Wort für Wort und prüft, ob das aktuelle Wort in einer der beiden zuvor erstellten Listen von Krankheiten und Symptomen auftritt. Falls ja, wird in der artikelspezifischen CSV-Datei ebenfalls nach dem Wort gesucht. Taucht es bereits in der CSV-Datei auf, wird der Wert des Auftretens um eins erhöht, ansonsten wird ein neuer Eintrag mit dem Wert eins angelegt. Am Ende dieses Prozesses liegen für jeden Artikel zwei CSV-Dateien vor, eine für die Krankheiten und eine für die Symptome. Darin sind alle in den Abstracts der Artikel auftauchenden Krankheiten beziehungsweise Symptome sowie ihre Auftrittshäufigkeit aufgelistet.

Um die Daten für spätere Visualisierungen vorzubereiten, werden die Häufigkeiten nicht nur auf Artikelebene, sondern auch für verschiedene Zeitintervalle abgebildet. Ausgewählt werden hierbei zum einen die einzelnen Jahre und zum anderen die drei Zeiträume vor 2000, 2000–2009, 2010–2020. Die Zeiträume sind dabei bewusst gewählt, da seit dem Jahr 2000 ein deutlicher Zuwachs an Artikeln in dem Datenkorpus zu erkennen ist und so eine ungefähr gleichmäßige Verteilung gewährleistet werden kann.

Beim Zusammenfassen der im vorherigen Abschnitt erstellten CSV-Datei kann auf das Namensschema dieser zurückgegriffen werden. Dieses wurde so gewählt, dass jede Datei den Jahreswert aus dem XML-Baumelement PubDate des jeweiligen Artikels enthält. PubDate gibt das Publikationsdatum des Artikels an. Für die einzelnen Jahre muss daher über alle Dateien iteriert werden und der Inhalt aus allen Dateien, die das aktuell gesuchte Jahr im Namen enthalten, in eine neue CSV-Datei geschrieben werden. Auch hier wird für jeden Krankheitsbeziehungsweise Symptomnamen erst geprüft, ob dieser bereits in der neuen CSV-Datei vorhanden ist. Falls nein, wird ein neuer Eintrag mit dem Wert aus der aktuell ausgelesenen CSV erstellt, falls ja, wird dieser Wert auf den bereits hinterlegten Wert addiert. Für die Zeiträume ist derselbe Prozess zu durchlaufen, nur, dass diesmal darauf geprüft wird, in welchen Zeitraum die aktuelle CSV-Datei (Jahresebene) fällt. Die Zeiträume werden anhand der Grenzen Jahreszahl < 2000, $2000 \leq$ Jahreszahl < 2010 und $2010 \leq$ Jahreszahl ≤ 2020 durchgeführt.

Nach den ersten Ergebnissen der Visualisierungen wird ermittelt, welche Krankheiten beziehungsweise Symptome in wie vielen Artikeln genannt werden. Es wird demnach nur das Auftreten einer Krankheit/eines Symptoms in einem Artikel

betrachtet und nicht, wie häufig dieses in dem Artikel auftritt. Hierdurch soll verhindert werden, dass ein Artikel, in dem die Krankheit/das Symptom sehr häufig genannt wird, den Eindruck vermittelt, dass diese Krankheit/das Symptom insgesamt sehr häufig genannt wurde.

Am Ende der Zusammenfassung liegen sowohl für jedes Jahr als auch für jeden der definierten Zeiträume jeweils zwei CSV-Dateien für die Krankheiten und eine für die Symptome vor.

Im letzten Schritt werden Pärchen von Krankheiten beziehungsweise Symptome ermittelt, die häufig gemeinsam auftreten. Hierbei fiel die Entscheidung, nicht wie zuvor auf Jahre oder Zeiträume aufzuteilen, sondern den gesamten Korpus zusammenzufassen. Außerdem wird davon ausgegangen, dass eine einmalige gemeinsame Nennung in einem Abstract bereits als gemeinsames Auftreten genügt. Ein mehrfaches Auftreten der einzelnen Kombinationen innerhalb eines Abstracts wird daher nicht mehrfach gezählt. Auch hier wird wieder auf die erstellten CSV-Dateien für jeden Artikel zurückgegriffen. Es wird über jede dieser CSV-Dateien iteriert und eine interne Liste von allen möglichen Tupeln (Kombinationen) erstellt. Diese werden wiederum in eine große CSV-Datei geschrieben, auch hier mit der vorherigen Prüfung, ob es dieses Tupel bereits gibt. Wie oben beschrieben, wird eine Mehrfachnennung einer Kombination in einem Abstract nicht weiter betrachtet, daher wird der Wert in der großen CSV-Datei erst mit jedem weiteren Auftreten in einem weiteren Abstract um eins erhöht. Die Tupel werden bereits in der internen Liste immer alphabetisch sortiert, sodass es hier nicht zu einer Dopplung der Tupel kommen kann, weil die Begriffe in unterschiedlicher Reihenfolge in den Abstracts auftauchen. Am Ende dieses Prozesses liegen zwei neue Dateien vor. In der einen werden alle auftretenden Krankheitskombinationen und ihre Häufigkeit aufgelistet, in der anderen alle auftretenden Symptomkombinationen mit ihrer Häufigkeit.

4.4 Ermittlung der Veröffentlichungen pro Jahr

Die in diesem Abschnitt folgenden Auswertungen werden mit Python in der Version 3.7.5 und den folgenden Bibliotheken durchgeführt: wordcloud, spaCy, textacy, scikit-learn und pandas.

Um die Anzahl der erschienenen Artikel pro Jahr im Korpus auszuwerten, werden alle XML-Dateien der Artikel mit der Python-Methode os.walk() in einer Schleife durchlaufen. Für die Bestimmung des Jahres in den XML-Dateien wird das Publikationsjahr verwendet. Um die Beziehung zwischen dem Jahr und der Anzahl

der erschienenen Artikel zu erfassen, wird ein Dictionary verwendet. Als Schlüssel für das Dictionary wird das Erscheinungsjahr genutzt und für den Wert die Anzahl der erschienenen Artikel. Zum Aufbau des Dictionaries wird für jeden Artikel geprüft, ob das Erscheinungsjahr bereits erfasst wurde. Wenn das Jahr schon erfasst wurde, wird der dahinterliegende Wert um eins inkrementiert. Sollte das Jahr noch nicht erfasst sein, wird ein neuer Schlüssel mit dem Wert eins angelegt. Zur vereinfachten Darstellung wird das Dictionary aufsteigend nach den Schlüsseln sortiert.

Damit die Ergebnisse nicht verloren gehen, werden sie als CSV-Datei exportiert. Anhand dieser Datei kann das Ergebnis in einem vertikalen Balkendiagramm visualisiert werden. Für die Analyse werden fast ausschließlich die Python-Standardwerkzeuge verwendet. Zum Speichern der CSV-Datei kam die Bibliothek pandas zum Einsatz.

Das Diagramm in Abbildung 8 zeigt das Ergebnis dieser Untersuchung. Zu sehen ist, dass ab dem Jahr 1994 die Anzahl der veröffentlichten Artikel stark ansteigt und erst im Jahr 2013 abflacht und stagniert. Zum Zeitpunkt der Korpus-Erstellung im Herbst 2019 gab es ein Paper, dessen Veröffentlichung erst für das Jahr 2020 geplant war.

Abbildung 8: Publikationen pro Jahr

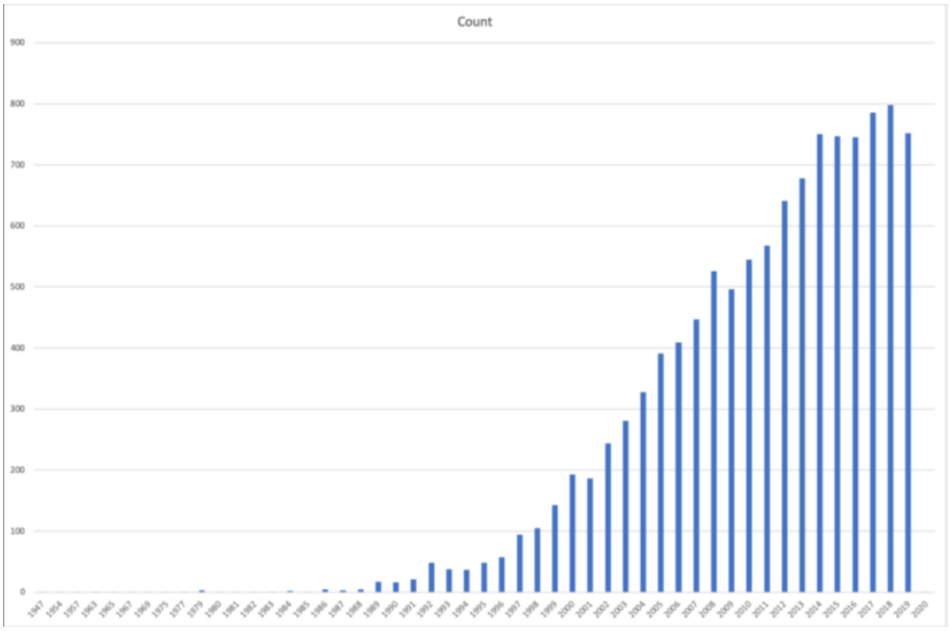
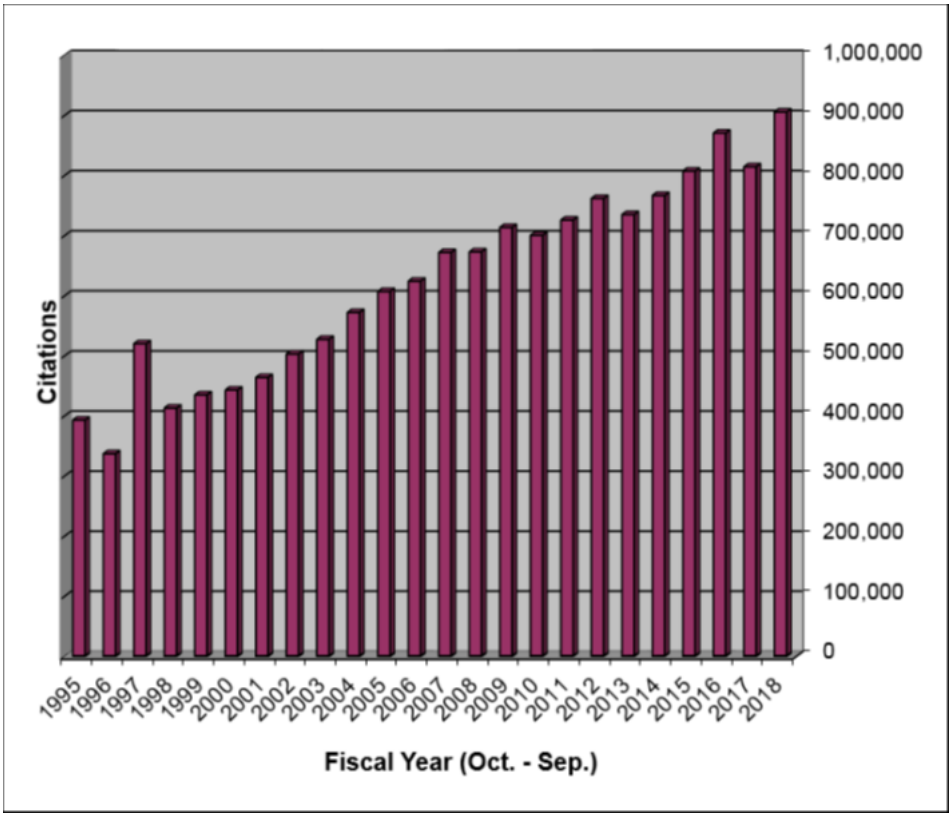


Abbildung 9: Entwicklung der Veröffentlichung in MEDLINE



Im Vergleich dazu ist in Abbildung 9 die Entwicklung der Veröffentlichung in MEDLINE insgesamt abgebildet. Hieraus lässt sich erkennen, dass das Interesse an Grüntee im Vergleich zu den Veröffentlichungen bei MEDLINE insgesamt überproportional ansteigt.

Zur Erstellung einer Wordcloud, welche die in dem Korpus enthaltenen Wörter anhand ihrer Häufigkeit visualisiert, ist es erforderlich eine Zeichenkette mit allen enthaltenen Wörtern zu erstellen. Dazu werden sie in eine Liste geladen, mit welcher dann ein textacy-Korpus erstellt werden kann. textacy ist eine Python-Bibliothek zum Vorbereiten einer Text-Analyse (vgl. Textacy, 2019), welche mit spaCy durchgeführt werden kann. spaCy wiederrum ist eine sehr schnelle NLP-Bibliothek, Natural Language Processing (vgl. SpaCy 2020), die die bearbeiteten Abstracts in sogenannte Docs unterteilt, welche sich wiederrum in Tokens und Spans gliedern. Für die Erstellung des Korpus wird jeder Abstract mit dem spaCy-

Modell für die englische Sprache analysiert und in einer Liste zwischengespeichert. Nachdem spaCy alle Abstracts bearbeitet und als Doc zurückgegeben hat, wird aus der Liste der Docs der textacy-Korpus erstellt und für weitere Analysen gespeichert.

Zum Erstellen der Wordcloud kann nun über alle in den Docs enthaltenen Tokens iteriert werden. Damit nicht alle Wörter ihren Weg in die Wordcloud finden, werden die sogenannten Stoppwörter während der Erstellung gefiltert. Um die Wörter auf ihren Stamm zurückzuführen, wird die von spaCy bereitgestellte lemmatisierte Zeichenkette des Tokens verwendet. Diese Wörter werden dann zu der Zeichenkette für die Wordcloud zusammengeführt. Die Python-Bibliothek wordcloud erzeugt anschließend aus der bereitgestellten Kette die Abbildung 10 und Abbildung 11.

Abbildung 10: Wordcloud mit Suchbegriffen

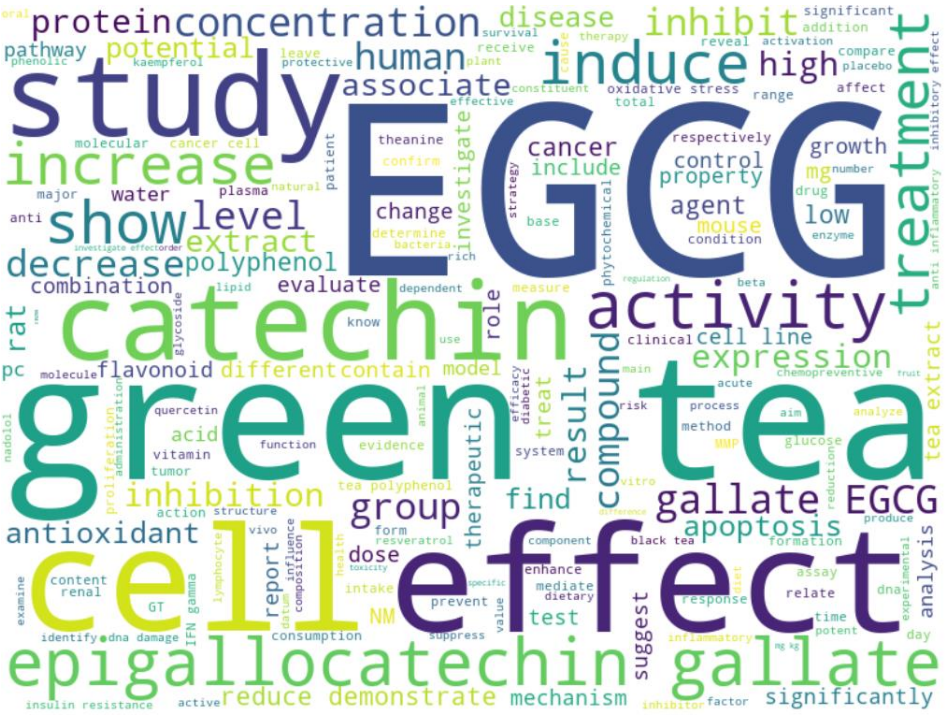
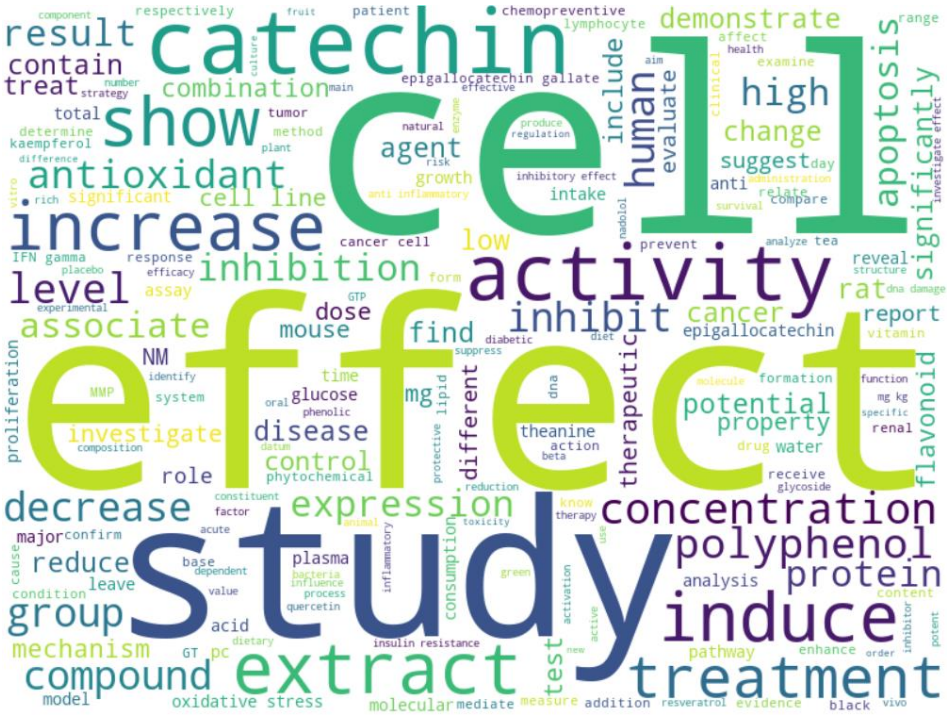


Abbildung 10 zeigt die am häufigsten vorkommenden Wörter in den analysierten Abstracts. Zu erkennen ist, dass die Suchbegriffe, welche zur Erstellung des Korpus verwendet wurden, am häufigsten vorkommen. Daher wird entschieden, eine weitere Wordcloud ohne die Suchbegriffe zu erstellen, wie in Abbildung 11 zu sehen ist.

Abbildung 11: Wordcloud ohne Suchbegriffe



Die verfeinerte Wordcloud lenkt den Fokus weg von den Suchbegriffen und hin zu den dahinterliegenden und am häufigsten vorkommenden Begriffen. Im Zusammenhang mit Grüntee respektive dem Wirkstoff EGCG kommt häufig das Wort cell vor. Des Weiteren ist festzustellen, dass die Wörter effect und catechin häufiger auftreten. Es ist also anzunehmen, dass ECGC auf der Zellebene wirkt. Aufgrund des Wortes treatment könnte abgeleitet werden, dass einige Artikel des Korpus EGCG als Wirkstoff in einer medizinischen Therapie untersuchen. Eine genauere Untersuchung bestätigt, dass grüner Tee (oder EGCG) tatsächlich häufig eingesetzt wird, um die Wirkung auf Krebszellen zu untersuchen. Ein markantes Beispiel ist die Untersuchung von Calgarotto et al., welche eine Wirksamkeit bei der Behandlung von Leukämie-Zellen nachweisen konnten (vgl. Calgarotto et al., 2018). Auffällig ist ebenfalls der Begriff „apoptosis“. Tatsächlich konnte nachgewiesen werden, dass grüner Tee in der Lage ist, den programmierten Zelltod (Apoptose) von Brustkrebszellen einzuleiten (vgl. Thangapazham et al., 2007). Beim genaueren Betrachten der Grafik sind weitere Substanzen in der Cloud erkennbar: Quercetin und Reservatrol sind Substanzen aus der

Gruppe der natürlich vorkommenden Polyphenole mit einem ähnlichen Wirkmechanismus wie EGCG. Interessant ist noch die Substanz Theanin – es handelt sich hier ebenfalls um einen Bestandteil von grünem Tee – L-Theanin ist eine Aminosäure, welche für ihre stimmungsaufhellende Wirkung bekannt ist (vgl. Lopes Sakamoto et al., 2019).

4.5 Analyse der Abstracts mit BERT

Ziel dieses Abschnitts ist die Prüfung, inwieweit die Technik BERT in der Lage ist, Fachartikeln zum Thema Green Tea entsprechende MeSH-Codes zuzuordnen. Bisher müssen zu jedem Artikel per Hand die thematisch korrekten Codes nachgetragen werden. Aufgrund der beschränkten Anzahl von Artikeln, der enorm hohen Menge an unterschiedlichen MeSH-Codes, sowie der Komplexität der Texte, wird im Folgenden versucht, lediglich konkrete Fragestellungen zu beantworten.

BERT ist eine von Google entwickelte Technik im Bereich des NLP und steht für Bidirectional Encoder Representations from Transformers und wurde 2018 als Open-Source-Modell öffentlich zur Verfügung gestellt (vgl. Devlin & Chang, 2018). Im Gegensatz zu anderen NLP-Modellen ist das Besondere an BERT, dass es kontextbezogen arbeitet. Bei anderen NLP-Modellen werden Stoppwörter häufig entfernt, um kontextfrei die Texte zu analysieren. BERT hingegen nutzt diese Stoppwörter bewusst, um bidirektional den Kontext erschließen zu können und somit optimierte Ergebnisse zu gewährleisten (vgl. Devlin et al., 2018). Zwar haben Stoppwörter keinen semantischen Mehrwert, BERT jedoch benötigt diese Informationen für die Analysen.

Die Effektivität von BERT bei der Textanalyse hat verschiedene Ansätze. Vor allem das Pre-Training ist bei der Analyse von Texten durch NLP-Modelle sehr zeit- und rechenintensiv. BERT stellt verschiedene Modelle zur Verfügung, die bereits vortrainiert sind. Somit kann mit vergleichsweise wenigen Daten, sehr schnell der spezifische Use-Case trainiert werden, womit viel Zeit und Ressourcen eingespart werden können. BERT nennt dies FineTuning. Das Modell BERT-Large besitzt beispielsweise 24 Layers mit 1.024 hidden Nodes und insgesamt 340 Millionen Parametern (vgl. BERT GitHub, 2019).

Da eine Multi-Label-Textklassifikation mit den entsprechenden Daten keine verwertbaren Ergebnisse lieferte, wird eine Multi-Class-Klassifikation durchgeführt.

Vermutlich war bei der Multi-Label-Klassifikation das Verhältnis zwischen Datensätzen und Label-Ausprägungen ein Problem und somit fiel die Entscheidung darauf, weniger komplexe Ausprägungen zu analysieren.

Deshalb wird der DescriptorName der MeSH-Headings betrachtet und es fällt auf, dass die Aufteilung dort unterschiedlich ist. Auf höherer Ebene der MeSH-Codes zeigt sich, dass diese animals, humans, beide oder keine der Ausprägungen enthalten. Aufgrund dessen wird ausgewertet, wie sich die Verteilung über alle XML-Dateien darstellt und es zeigt sich, dass diese vergleichsweise ausgeglichen verteilt sind. Die Anzahl der Artikel je Kategorie kann in folgender Abbildung betrachtet werden:

Abbildung 12: Verteilung Rohdaten auf Klassifikationen

```
Anzahl Artikel mit Mensch und Tier:  1128
Anzahl Artikel mit Tier:  1831
Anzahl Artikel mit Mensch:  2424
Anzahl Artikel ohne Mensch und Tier:  1080
```

Daraufhin wird ein Datensatz mit den Abstracts der Artikel und der dazugehörigen Zuordnung der verschiedenen Klassen erstellt, anschließend erfolgt eine Aufteilung in 80 Prozent Trainingsdaten und 20 Prozent Testdaten. Zum FineTuning des BERT-Modells wird auf die Bibliothek ktrain zurückgegriffen (vgl. Google Research, 2019). Diese nutzt im Hintergrund die Keras-Architektur und verwendet BERT als Modell.

In Abbildung 13 ist die Erzeugung des preproc zu erkennen, welches als Eingabe für das BERT-Training dient. Hierzu werden jeweils die Sets der Trainingsdaten übergeben sowie die Klassen, welche als Anhaltspunkte für das Training dienen sollen. Die Klassen werden mit dem Parameter class_names übergeben.

Abbildung 13: Initialisieren des preproc

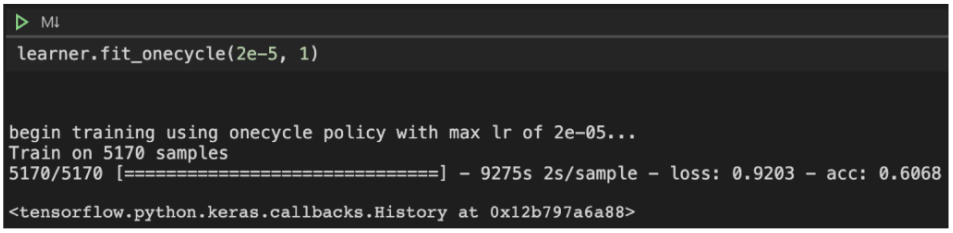
```
(x_train,  y_train), (x_test, y_test), preproc = text.texts_from_array(x_train=x_train, y_train=y_train,
                                                                       x_test=x_test, y_test=y_test,
                                                                       class_names=categorien,
                                                                       preprocess_mode='bert',
                                                                       ngram_range=1,
                                                                       maxlen=256,
                                                                       max_features=35000)
```

Im nächsten Schritt muss das zu trainierende Modell instanziiert werden. Dies wird mit dem zuvor erstellten preproc erzeugt.

Abbildung 14: Erzeugung des Multi-Class-Modells

Da das Modell nun erzeugt ist, kann dieses nun mit den Trainingsdaten trainiert werden. Dazu wird zunächst ein Learner erzeugt, der danach entweder in mehreren Zyklen oder auch nur in einem Lernzyklus trainiert werden kann.

Abbildung 15: Training des Modells



Die Abbildung zeigt, dass mit insgesamt 5.170 Beispielen trainiert und eine Genauigkeit von ca. 60 Prozent während des Trainings erzielt wird. Dabei wurde das Modell jedoch nur mit der Durchführung einer Epoche trainiert. Aus diesem Grund wird BERT noch einmal mit drei Epochen trainiert und erlangt eine Genauigkeit von ca. 82 Prozent, allerdings nur auf den bereits bekannten Trainingsdaten. Daraufhin wird dieses trainierte Modell mit den Testdaten validiert, wobei die Testdaten Datensätze enthalten, die dem Modell unbekannt sind. Abschließend lässt sich festhalten, dass bei der Validierung des Modells eine Genauigkeit von 74 Prozent erzielt wird.

Abbildung 16: Validierung des Modells

```
M↓
learner.validate(val_data=(x_test, y_test))
```

	precision	recall	f1-score	support
0	0.52	0.47	0.49	200
1	0.83	0.75	0.79	392
2	0.76	0.79	0.77	471
3	0.74	0.88	0.80	230
accuracy			0.74	1293
macro avg	0.71	0.72	0.72	1293
weighted avg	0.74	0.74	0.74	1293

Das Ergebnis verdeutlicht, dass BERT nur mit Hilfe des Abstracts in der Lage ist, mit einer Genauigkeit von 74 Prozent zu bestimmen, ob einem Artikel die MeSH-Headings humans und animals, beide oder keines von beiden zugeordnet werden können.

5 Ergebnisse und Visualisierung

Nachdem die Daten zu einem Korpus zusammengefügt und zur Analyse vorbereitet wurden, können diese Daten mit Big Data-Analyse Tools weiterbearbeitet werden. In diesem Kapitel werden zunächst die MeSH-Codes näher untersucht, bevor der Fokus auf die konkreten Wörter der Artikel gelegt wird. Zuletzt werden explizit die aufkommenden Krankheiten und Symptome sowie die Publikationsdaten untersucht.

5.1 Verwendete Tools zur Visualisierung

Power BI von Microsoft stellt eine Sammlung von Business-Intelligence-Tools für die cloudbasierte Analyse von Geschäftsdaten dar. Die Lösung setzt sich aus mehreren Komponenten zusammen und besteht aus den zentralen Power BI-Services und damit verbundenen Benutzeroberflächen, Datenbanken und Gateways für die Anbindung verschiedener Datenquellen. Mithilfe von Power BI lassen sich Geschäftsdaten analysieren und die Ergebnisse grafisch aufbereiten.

Das BI-Tool Tableau Desktop ist ein Business-Intelligence- und Analytics-Softwareprodukt der US-amerikanischen Firma Tableau Software. Mit Hilfe dieses Produktes können Daten aus unterschiedlichen Quellen in ein Modell geladen, mit weiteren Daten aggregiert und visualisiert werden. Als Datenquelle können verschiedene Datenaustauschformate wie CSV, Microsoft (MS) Excelformat, PDF und viele weitere genutzt werden. Die eingelesenen Daten können bei Tableau Desktop in einem begrenzten Umfang bearbeitet werden. Unter anderem ist es möglich, den Datentyp von Tabellenspalten zu ändern, Tabellenspalten auszublenden und Beziehungen zwischen unterschiedlichen Datenquellen automatisch und manuell zu bearbeiten. Tableau Desktop ist ein kostenpflichtiges Produkt, welches Studierenden jedoch zeitlich begrenzt kostenlos zur Verfügung gestellt wird. Es eignet sich zur schnellen und einfachen Datenmodell-Erstellung und Visualisierung. Ausführliche Analysen und Vorhersagen unter Zuhilfenahme von Algorithmen des maschinellen Lernens sollten auf Grund von Limitationen außerhalb von Tableau durchgeführt werden.

MicroStrategy ist eine Firma, die ein gleichnamiges Produkt entwickelt hat. Bei dem Produkt handelt es sich um eine Sammlung von BI-Tools, welche verschiedene Techniken und Anwendungsbereiche nutzen, um Daten zu analysieren, Geschäftsprozesse zu verbessern bzw. zu beschleunigen und Unternehmen damit bei der Digitalisierung zu helfen. Für diese Ausarbeitung wurde MicroStrategy

Analytics genutzt, um eine explorative Datenanalyse durchzuführen. Die Entscheidung fiel auf dieses Tool, da es unter anderem ermöglicht, sowohl einfache als auch komplexe Visualisierungen zu erstellen. Nach einem Datenimport können verschiedene Visualisierungstypen wie Boxplots, Streudiagramme und Balkendiagramme mit wenigen Klicks erstellt werden. Zusätzlich beinhaltet das Tool die Möglichkeit, dynamische Filter zu nutzen.

Gephi ist ein quelloffenes Softwarepaket zur Netzwerkanalyse und Visualisierung, das in der Programmiersprache Java geschrieben wurde, und verwendet die Plattform NetBeans. Es eignet sich besonders gut zur Darstellung und Verbindung von Entitäten. Gephi verwendet zur Visualisierung OpenGL und damit die Grafikkarte zur graphischen Darstellung. Das erlaubt es, Gephi-Netzwerke mit über 20.000 Knoten zu verarbeiten, ohne dabei den Haupt-Prozessor zu beanspruchen.

5.2 Exploration der MeSH-Codes

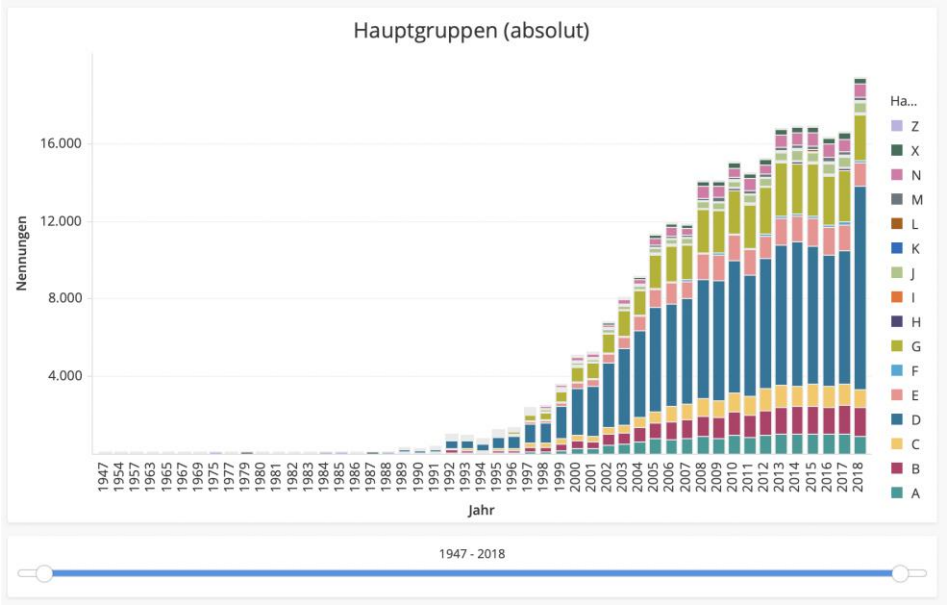
Zunächst einmal wurde auf Basis des Korpus eine Exploration der MeSH-Codes mit Fokus auf Hauptgruppen durchgeführt und visualisiert. Diese Visualisierung wurde nach Zeiträumen differenziert, also eine Längsschnittbetrachtung vollzogen, in der eine Entwicklung der Themen im Zeitverlauf sichtbar wird. Eine Querschnittsbetrachtung, die das Gesamtaufkommen aller MeSH-Kategorien über alle Zeiträume zeigt, vervollständigt das Bild auf der ersten Ebene. In weiteren Schritten werden die MeSH-Codes weiter heruntergebrochen und die Subgruppen bis zur fünften Ebene visualisiert. Dies geschieht auf verschiedene Weise. Einerseits werden die interessantesten Kategorien wie *Diseases* (C), *Anatomy* (A) und *Chemicals and Drugs* (D) in verschiedenen Visualisierungsformaten näher betrachtet. Zum Abschluss werden die MeSH-Code-Paarungen auf den Ebenen 3 bis 5 gebildet, um gemeinsame Auftrittshäufigkeiten darzustellen.

Für den ersten Schritt wurde der Datensatz in MicroStrategy importiert. Um nun eine Analyse der Hauptgruppen der MeSH-Codes durchzuführen, muss die entsprechende Information aus den MeSH-Codes extrahiert werden. Dazu wird das erste Zeichen eines MeSH-Codes ausgelesen und als Hauptgruppe gespeichert. Insgesamt beinhaltet der Datensatz 16 Hauptgruppen, wozu unter anderem *Anatomy* (A), *Organisms* (B) und *Chemicals and Drugs* (D) zählen. Der Datensatz enthält außerdem pro MeSH-Code die Nennungen pro Jahr.

Zusätzlich wurde aufgrund der Menge an Granularität der MeSH-Codes eine relationale Datenbank mit der Datenbank-Software MariaDB erstellt, um Aggregationen und Filterungen des Korpus, basierend auf den Metainformationen, einfach durchführen zu können. In den nachfolgenden Analysen wurde daher immer auf die Datenbank zurückgegriffen, um kleinere Datensets oder Aggregationen abzufragen. Für die Integration des Korpus in die Datenbank wurde das ETL-Tool Talend Data Integration Studio verwendet. Für die weitere Verwendung in den Visualisierungstools Tableau oder Gephi wurden die Daten entweder über den SQL-Client DBeaver abgefragt und anschließend als Dateiexport im CSV-Format bereitgestellt oder über Python-Code mit den Bibliotheken Numpy, Pandas und Networkx gearbeitet.

5.2.1 MeSH-Codes im Zeitverlauf

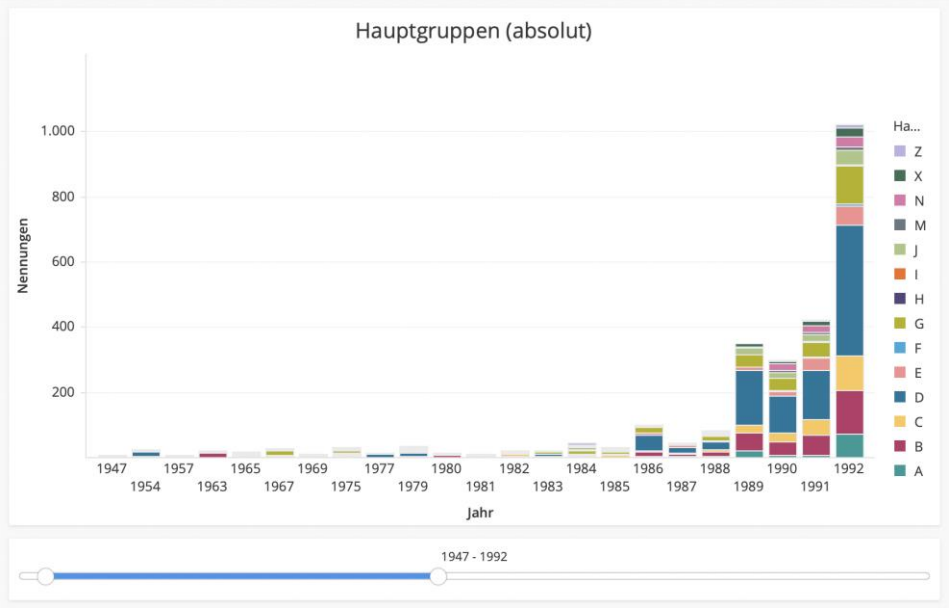
Abbildung 17: Anzahl der Nennungen pro Hauptgruppe pro Jahr (bereinigt)



Nach Bereinigung der zu ignorierenden Daten fällt auf, dass die Nennungen seit den 1990er Jahren rasant gestiegen sind. Aufgrund des großen Unterschieds in der Anzahl der Nennungen über den gesamten Zeitraum lassen sich mit der hier gewählten Skala die Jahre vor den 1990ern nur schlecht erkennen. Zwischen den Jahren 1986 und 1989 entwickelt sich eine relativ feste Verteilung, welche

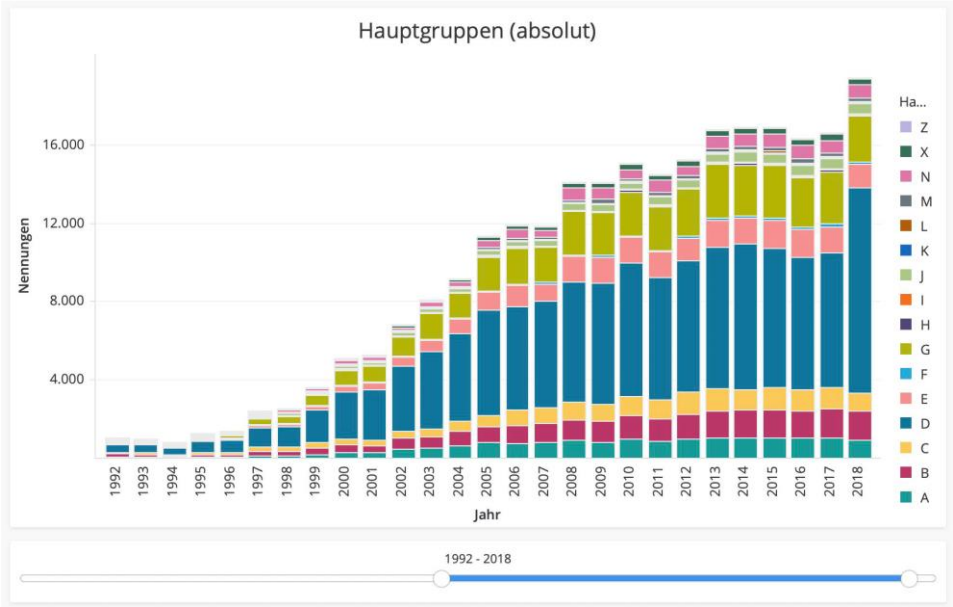
sich bis heute recht gleichmäßig weiterentwickelt. Um die gewählte Skala aufzuteilen, wird die Visualisierung im nächsten Schritt in zwei Teile getrennt, zum einen von 1947 bis 1992 und zum anderen von 1992 bis 2019.

Abbildung 18: Anzahl der absoluten Nennungen pro Hauptgruppe pro Jahr (1947–1992)



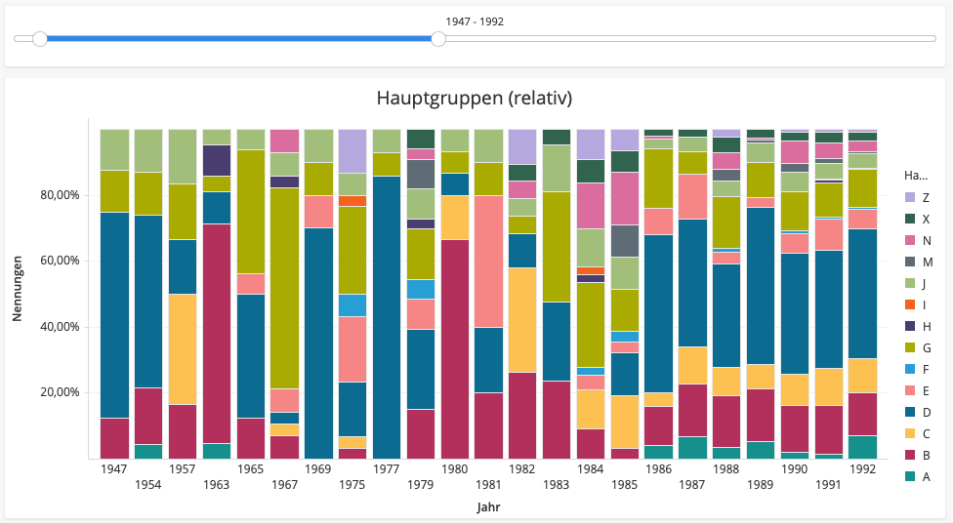
Anhand von Abbildung 18 wird deutlich, dass auch bei neu gewählter Skala vor dem Jahr 1986 im Vergleich kaum Nennungen vorhanden sind. Im Jahr 1989 steigt die Anzahl der Nennungen zum ersten Mal rasant auf über 300 an. Der nächste große Anstieg von Nennungen findet im Jahr 1992 statt. Hier werden über 1.000 Nennungen verzeichnet, mehr als doppelt so viele wie im vorherigen Jahr.

Abbildung 19: Anzahl der absoluten Nennungen pro Jahr (1992–2018)



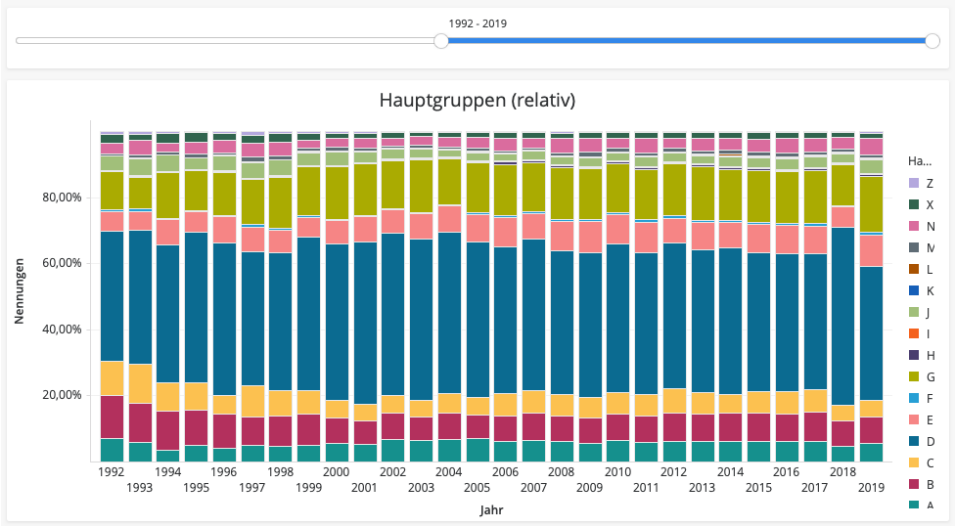
Der in Abbildung 19 gewählte Zeitraum zeigt, dass das zuvor betrachtete Jahr 1992 trotz des Anstiegs im Vergleich zu den folgenden Jahren nur eine geringe Anzahl an Nennungen aufweist. Die Hauptgruppe D (*Chemicals and Drugs*, hier als vierte von unten dunkelblau dargestellt) hat in den meisten Jahren die größte Anzahl an Nennungen. Zudem hat diese im Jahr 2018 im Vergleich zum Vorjahr einen deutlichen Anstieg erhalten. Um dieses Phänomen besser betrachten zu können, wird im nächsten Schritt die relative Häufigkeit jeder Hauptgruppe in den beiden Zeiträumen betrachtet.

Abbildung 20: Anzahl der relativen Nennungen pro Hauptgruppe pro Jahr (1947–1992)



Anhand von Abbildung 20 lässt sich erkennen, dass die Hauptgruppe D zwischen den Jahren 1947 bis einschließlich 1985 stark voneinander abweichende Anzahlen an Nennungen hatte. Aus diesem Grund ist hierbei kein klarer Trend erkennbar. Dies hat unter anderem damit zu tun, dass die Anzahl der absoluten Nennungen für diese Jahre sehr gering sind und dadurch eine enorme Schwankung der relativen Nennungen vorhanden ist. Ab dem Jahr 1986 ist eine konstante Steigung der Anzahl von Nennungen zu erkennen.

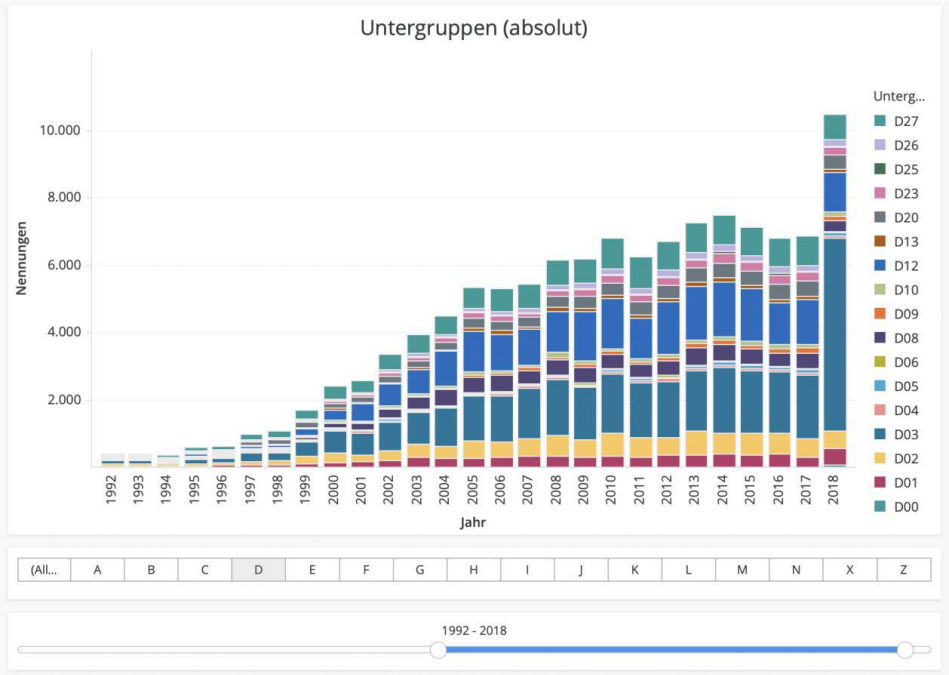
Abbildung 21: Anzahl der relativen Nennungen pro Hauptgruppe pro Jahr (1992–2019)



Durch Abbildung 21 kann der Anstieg im Jahr 2018 genauer betrachtet werden. Im Vergleich zu anderen Hauptgruppen hat die Hauptgruppe D in diesem Jahr verhältnismäßig zu den Gesamtnennungen stark zugelegt.

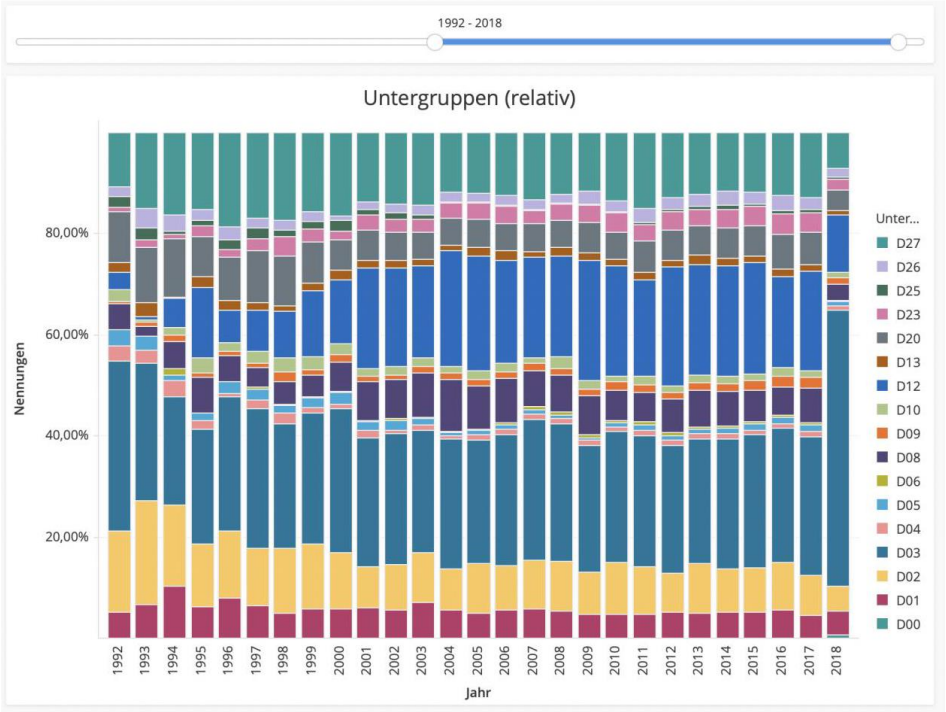
Um diesen Anstieg der Hauptgruppe D im Jahr 2018 genauer zu betrachten, wird im nächsten Schritt eine Aufteilung zu Untergruppen vorgenommen. Dazu werden die ersten drei Zeichen der MeSH-Codes extrahiert und als Untergruppe gespeichert. Auf diese Art und Weise lässt sich feststellen, welche Medikamente und Chemikalien häufig in Zusammenhang mit grünem Tee erwähnt werden.

Abbildung 22: Anzahl der absoluten Nennungen der Untergruppen von Hauptgruppe D



Die hier erstellte Visualisierung beinhaltet einen neuen dynamischen Filter. Dieser filtert die Daten nach Hauptgruppe D und zeigt deren Untergruppen für den angegebenen Zeitraum an. Im Hinblick auf den bereits festgestellten Anstieg der Nennungen von MeSH-Codes aus Gruppe D ist zu erkennen, dass die Anzahl der Nennungen der Untergruppe D03 (*Heterocyclic Compounds*) im Jahr 2018 stark gestiegen ist. Die Steigung von 1.879 Nennungen im Jahr 2017 auf 5.706 Meldungen im Jahr 2018 ist so extrem, dass die Untergruppe wahrscheinlich zu einem großen Teil für den Anstieg der Nennungen von Hauptgruppe D verantwortlich ist. Insgesamt ist die große Anzahl an Nennungen aus der Gruppe D03 nicht überraschend, da EGCG eine heterocyclische Substanz darstellt. Um einen zweiten Blickwinkel auf die Daten zu geben, wird die Visualisierung im nächsten Schritt auf eine relative Anzeige umgestellt.

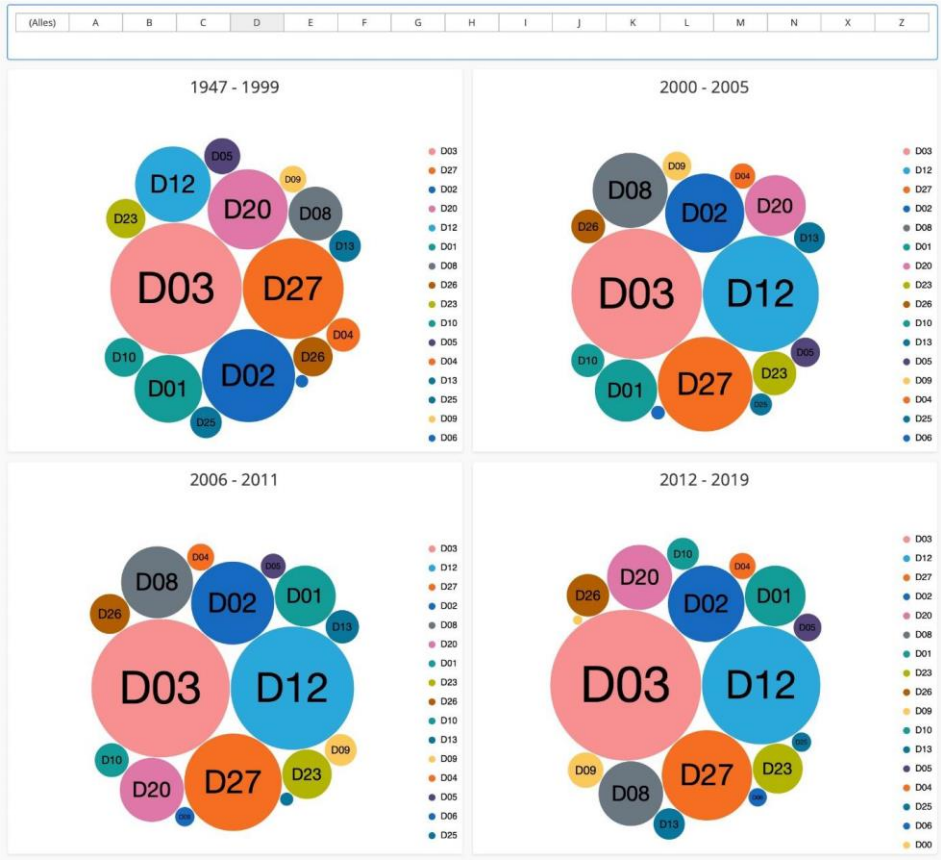
Abbildung 23: Anzahl der relativen Nennungen der Untergruppen von Hauptgruppe D



Auch in der relativen Ansicht ist gut zu erkennen, dass die Nennungen der Untergruppe D03 2018 im Verhältnis zu den vorherigen Jahren angestiegen sind. Aus diesem Grund haben alle weiteren Untergruppen in diesem Jahr prozentual an Anteil verloren. Auffällig ist das Auftreten der Codes aus der Rubrik D12 (*Amino Acids, Peptides, and Proteins*). Dieses kann darauf zurückführen, dass EGCG wohl häufig mit diesen Substanzklassen in Wechselwirkung tritt. Auch die relative Häufigkeit D08 (*Enzymes and Coenzymes*) spricht für eine Wechselwirkung mit Biokatalysatoren. Für die Wechselwirkung mit weiteren biologischen Substanzen spricht die Häufung der Gruppe D23 (*Biological Factors*).

Um den Effekt bezogen auf größere Zeiträume zu betrachten, werden im nächsten Schritt mehrere Circular Tree Maps erstellt und zum Vergleich nebeneinandergestellt.

Abbildung 24: Circular Tree Maps der Untergruppen von Hauptgruppe D in Zeiträumen



Hierbei wird deutlich, dass der zuletzt genannte Anstieg der Untergruppe D im Zusammenhang von größeren Zeiträumen schwächer ausfällt.

5.2.2 MeSH-Codes im Querschnitt

Im nächsten Schritt soll das Attribut Name als Gruppierungsmerkmal dienen, da die Ansicht der einzelnen Jahre ohne Visualisierung der Hauptgruppen wenig Aussagekraft hat. Dafür werden die Daten über den gesamten Zeitraum betrachtet.

Da wir nicht mehr nur die MeSH-Codes an sich betrachten und auch keinen Jahresverlauf darstellen, sondern nur die entsprechenden Bezeichnungen und deren

Auftreten im Detail betrachten möchten, konzentrieren wir uns bei der folgenden Analyse lediglich auf die Unique IDs, um doppelte Nennungen und verfälschte Aussagen zu vermeiden.

Nachdem die Entwicklung der MeSH-Codes im Zeitverlauf gezeigt wurde und eine Bewegung in den Subkategorien sichtbar war, wird in den nächsten Grafiken die gesamte Häufigkeitsverteilung der Publikationen nach MeSH-Codes visualisiert.

Abbildung 25: Verteilung der Publikationen auf MeSH-Codes-Hauptgruppen

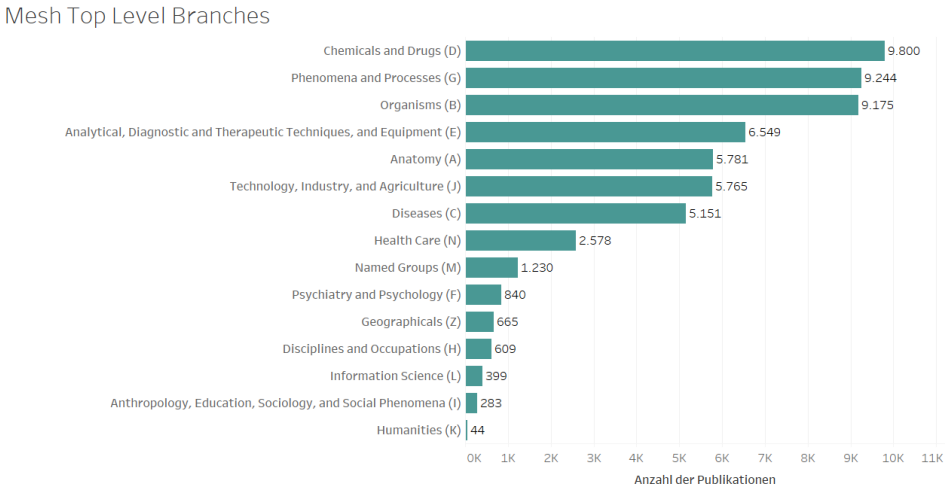
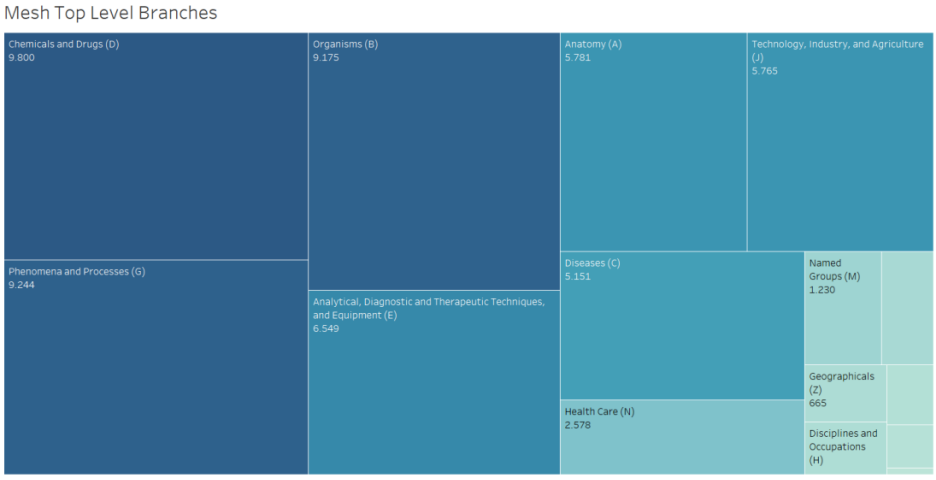


Abbildung 26: Verteilung der Publikationen auf MeSH-Codes-Hauptgruppen (Baumkasten-Struktur)



Die beiden Grafiken zeigen, wie stark die Hauptgruppen des MeSH-Baums in den Publikationen vertreten sind. Erkennbar ist zunächst das insgesamt starke Auftreten der Gruppen *Chemicals and Drugs* (D), *Phenomena and Processes* (G) und *Organism* (B) in 9.100 bis 9.800 Publikationen von insgesamt 11.605. Im Vergleich dazu ist erwähnenswert, dass *Diseases* (C) deutlich seltener betrachtet wird, nur 5.151 Publikationen sind mit MeSH-Codes zu Krankheiten versehen.

Abbildung 27: Verteilung der Publikationen auf die Top 10 MeSH-Codes (Ebene 2)

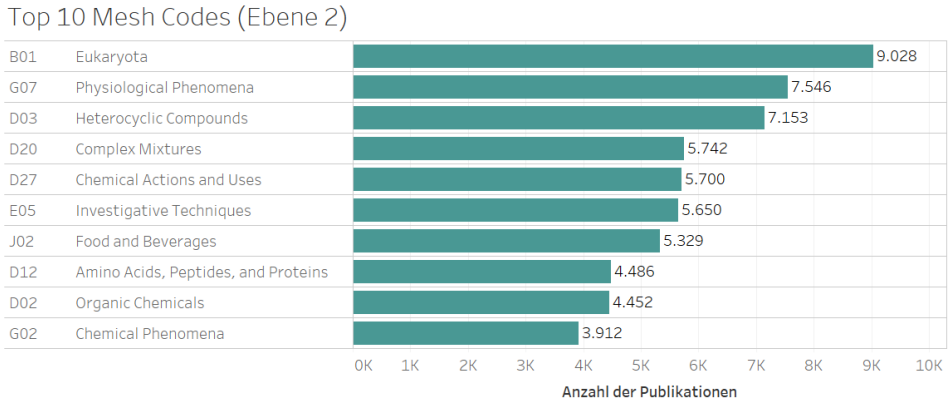


Abbildung 28: Verteilung der Publikationen auf die Top 10 MeSH-Codes (Ebene 1 und 2)

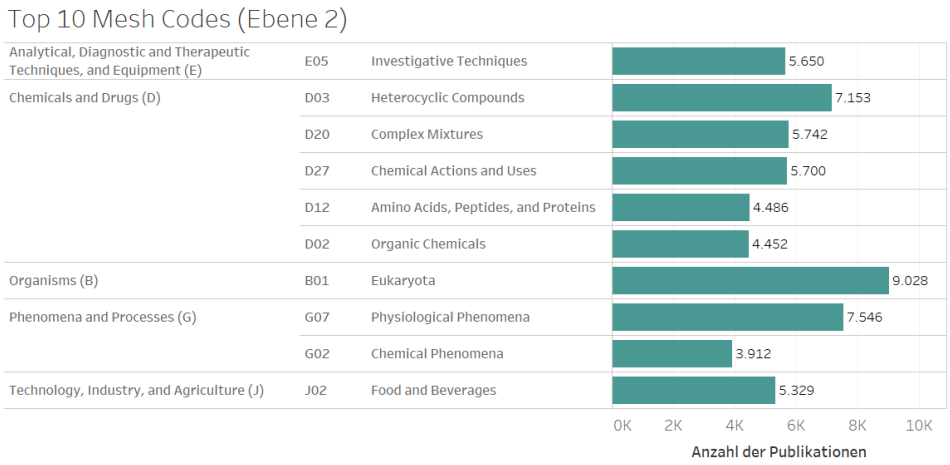


Abbildung 27 und Abbildung 28 zeigen die Verteilung der Publikationen nach dem MeSH-Code auf der zweiten Unterebene. Es ist ersichtlich, dass die aus den Hauptgruppen erkennbaren Häufungen bei B, G und D auch in den Top 3 MeSH-Codes auf der zweiten Ebene vertreten sind. So sind Eukaryota (B01) mit 9.028, Physiological Phenomena (G07) mit 7.546 und Heterocyclic Compounds (D03) mit 7.153 Publikationen am stärksten vertreten. Von den Top 10 der MeSH-Codes auf Ebene 2 sind fünf *Chemicals and Drugs* (D), zwei *Phenomena and Processes* (G) und jeweils ein Code *Organism* (B), *Analytical*, *Diagnostic and*

Therapeutic Techniques, *and Equipment* (E) und *Technology, Industry, and Agriculture* (J) zuzuordnen.

Abbildung 29: Verteilung der Publikationen auf die Top 10 MeSH-Codes (Ebene 3)

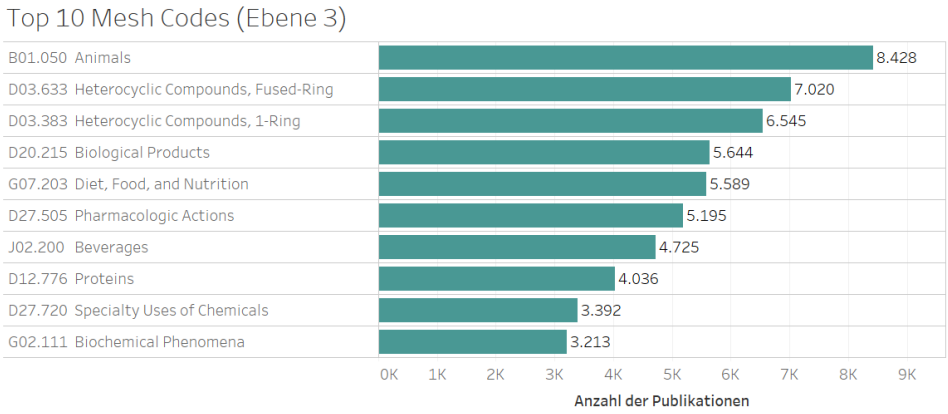
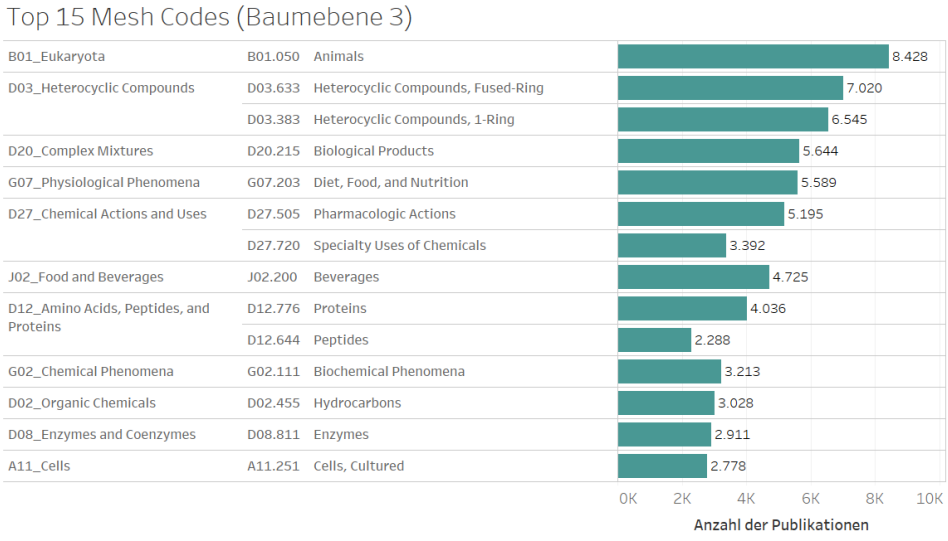


Abbildung 30: Verteilung der Publikationen auf die Top 15 MeSH-Codes (Ebene 2 und 3)

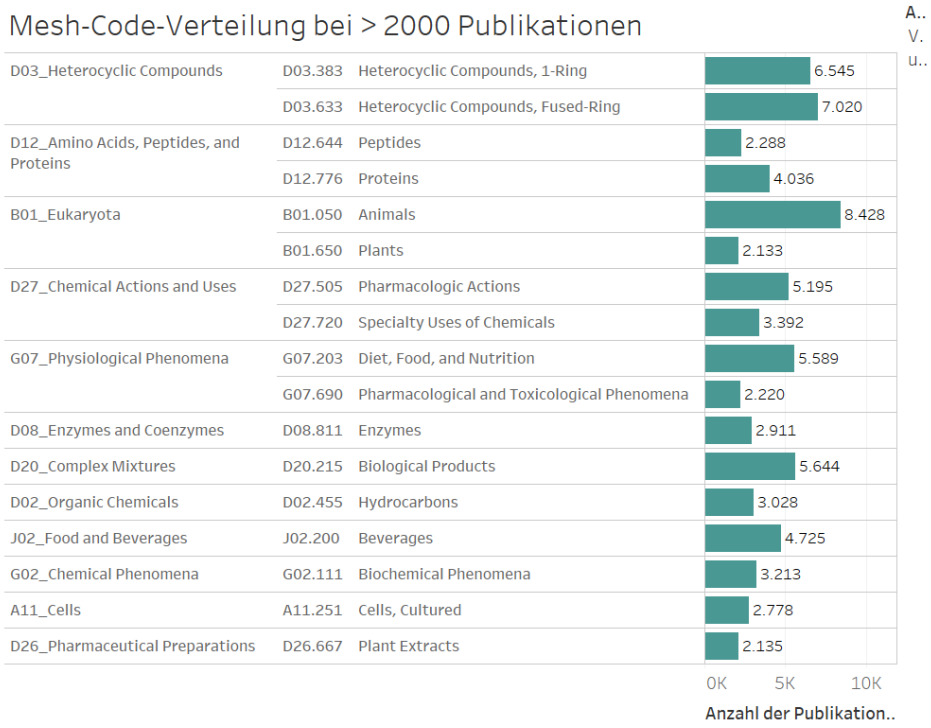


Die Grafiken zeigen die Verteilung der Publikationen nach den MeSH-Codes ab der dritten Unterebene an. Es ist erkennbar, dass die Hauptgruppen B, G und D in den Top 10 MeSH-Codes auf Ebene 3 vertreten sind. Wie auf Ebene 2, so ist

auch hier D mit den sechs Codes Heterocyclic Compounds, FusedRing (D003.633), Heterocyclic Compounds, 1-Ring (D03.383), Biological Products (D20.215), Pharmacologic Actions (D27.505), Proteins (D12.776), und Specialty Uses of Chemicals (D27.720) am stärksten vertreten. Darauf folgen zwei der Hauptgruppe G Diet, Food, and Nutrition (G07.203) und Biochemical Phenomena und jeweils ein zu B und J, Animals (B01.050) und Beverages (J02.200).

In der folgenden Visualisierung wird die Verteilung der MeSH-Codes auf Unterebene 3, gruppiert nach Ebene 2, mit mehr als 2.000 Publikationen sichtbar.

Abbildung 31: Verteilung der MeSH-Codes (Ebene 2 und 3) mit über 2.000 Publikationen

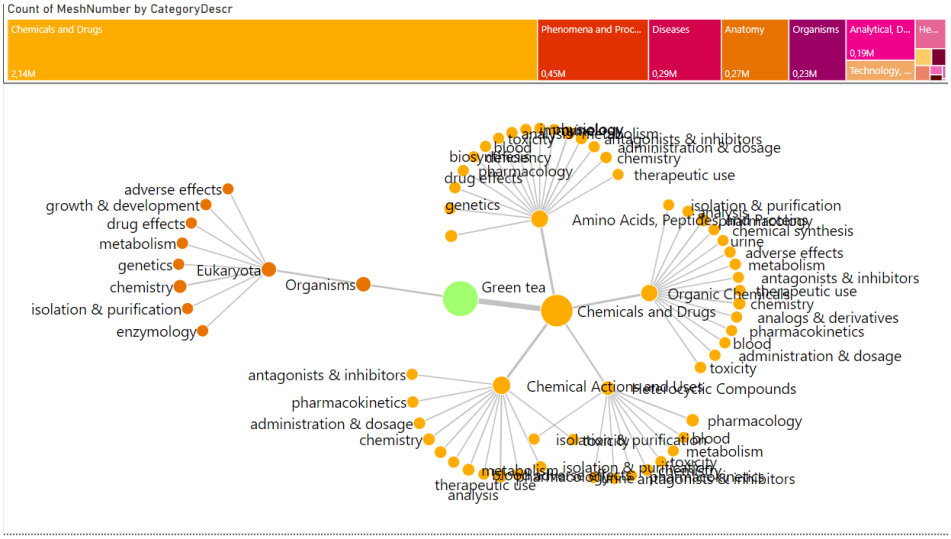


5.2.3 MeSH-Subgruppen

Besonders interessant ist, wie schon die vorangegangene Analyse in Abschnitt 5.2.1 zeigt, die Hauptgruppe D *Chemicals and Drugs*. Daher werden zunächst einige nähere Betrachtungen mittels verschiedener Darstellungen vollzogen.

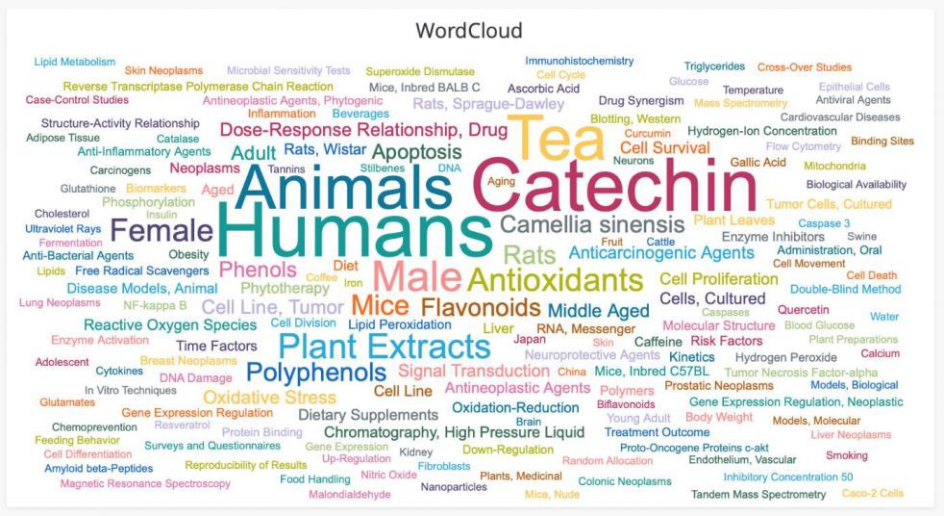
Als erstes wird eine Darstellung erzeugt, welche die Abhängigkeiten von *Chemicals and Drugs* und grünem Tee zu weiteren Begrifflichkeiten zeigt. Auch hier wird der vermehrte Einsatz von grünem Tee in der Chemie und Pharma-Industrie deutlich und weitere Bereiche, mit denen er konnotiert ist.

Abbildung 32: *Chemicals and Drugs* (D) (Power BI)



Bildet man für die Hauptgruppe D eine Wordcloud, so ergibt sich folgende Visualisierung.

Abbildung 33: Wordcloud über die Namen der MeSH-Unique-IDs



Hierbei ist zu erkennen, dass vier Namen mit einem relativ großen Abstand die größte Anzahl von Nennungen haben. Dazu gehören Humans (5.604), Catechin (5.442), Tea (4.541) und Animals (4.492). An fünfter Stelle steht Male mit 2.845 Nennungen.

Führt man Analysen auf der Hauptgruppe C *Diseases* durch, erhält man unter anderem folgende Darstellungen.

Abbildung 34: *Diseases* (C) (Power BI)

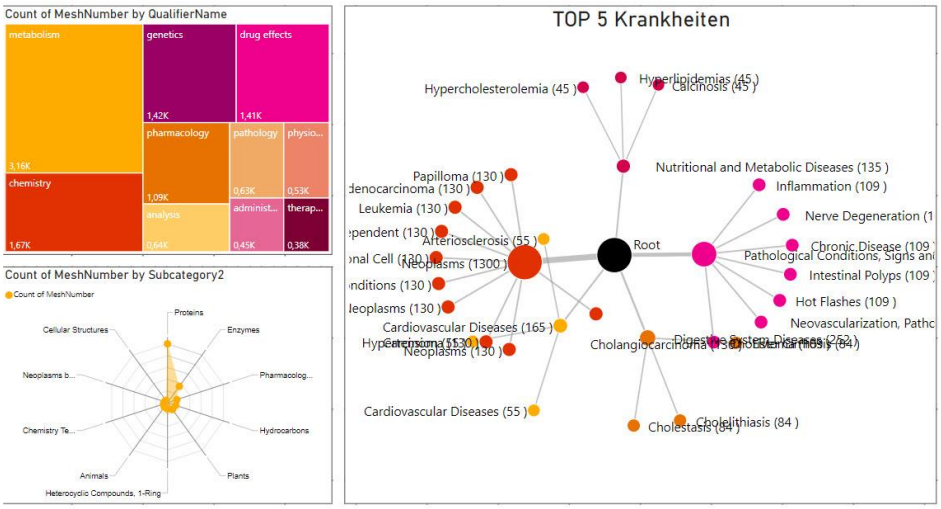


Abbildung 34 veranschaulicht die Top 5 Krankheiten, die laut der Anzahl der MeSH-Codes am stärksten in Relation zu grünem Tee stehen (Root). Das ist an erster Stelle Neoplasms (49,17 Prozent), gefolgt von Pathological Conditions, Signs and Symptoms (29,18 Prozent), Digestive System Diseases (9,64 Prozent), Cardiovascular Diseases (6,31 Prozent) und Nutritional and Metabolic Diseases (5,16 Prozent). In Bezug zu diesen Krankheiten wurden auch die stärksten Subgruppen ermittelt. Hierzu zählen hauptsächlich Proteins und Enzymes.

Abbildung 35: *Anatomy* (A) (Power BI)

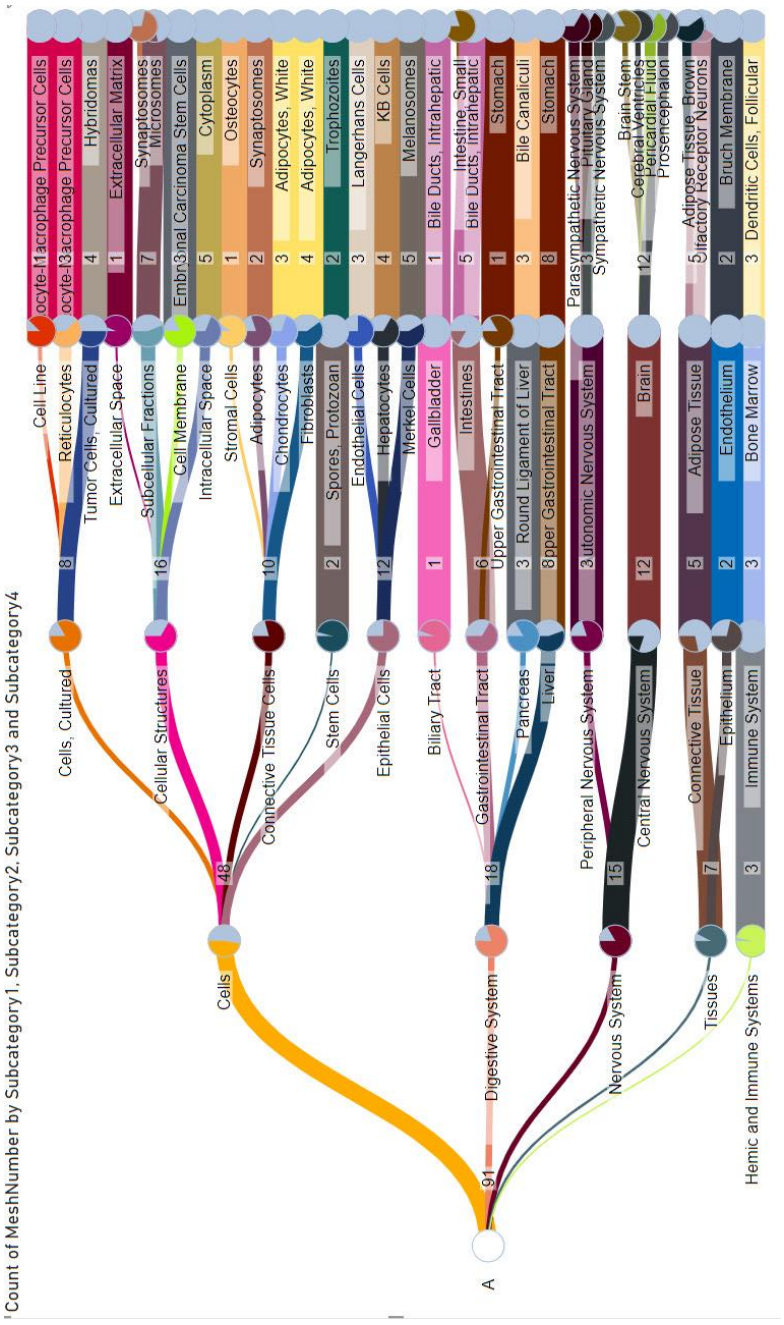


Abbildung 35 stellt eine Visualisierung der häufigsten MeSH-Codes in der Hauptgruppe A *Anatomy* dar, unterteilt nach den Top 5 Subgruppen, die die meisten MeSH-Codes beinhalten. Die Dicke der Verbindungslinien zwischen den verschiedenen Kategorien veranschaulicht die Anzahl der MeSH-Codes, die zu diesem Hierarchie-Element gehören. Die Kategorien mit den meisten MeSH-Codes sind Cells, Digestive System, Nervous System, Tissues und Hemic and Immune Systems. Überwiegend gehören die MeSH-Codes zu Cells (48 Prozent), wobei 16 Prozent davon wiederum zu Cellular Structures zählen. Auf der untersten Ebene wird deutlich, dass grüner Tee meist in Verbindung mit Synaptosomes, Microsomes und Cytoplasm gesetzt wird.

5.2.4 MeSH-Code-Paarungen

Die Datenbasis für die folgende Analyse bilden die Metainformationen zu den MeSH-Codes und deren Auftreten in den Publikationen. 10.003 von insgesamt 11.605 Publikationen sind mindestens einem MeSH-Code zugeordnet. Die Analyse wird auf diese 10.003 Publikationen beschränkt.

Die Selektion und Filterung erfolgten durch eine SQL-Abfrage in einem Python-Code an die vorher vorbereitete MariaDB-Datenbank. Ebenso wurden weitere Transformationen, wie die Kombinationsbildungen von MeSH-Codes in Python realisiert.

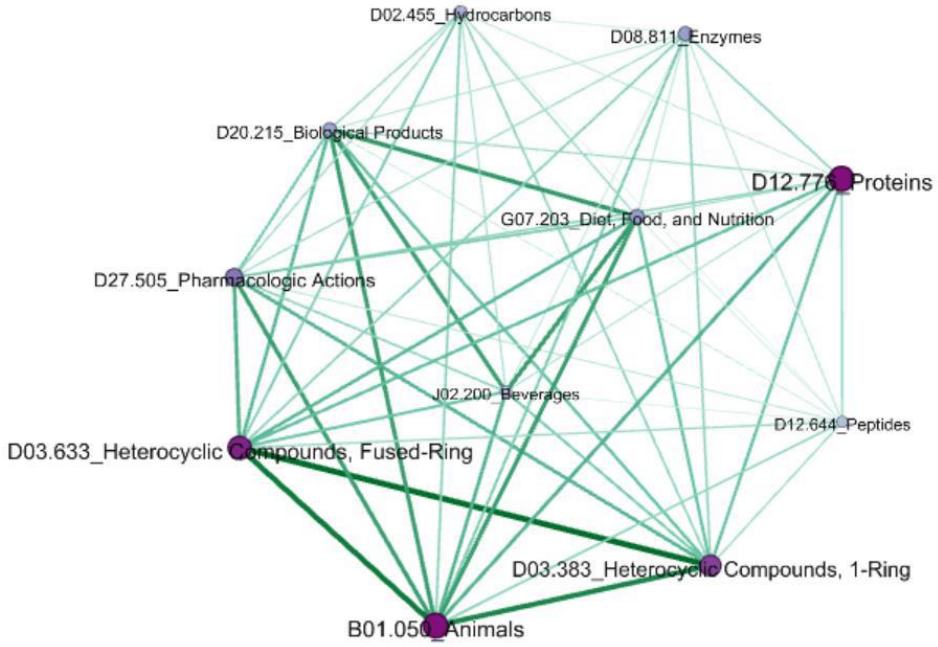
Es wurden auf den MeSH-Subgruppen 2 bis 5 jeweils eindeutige zweier MeSH-Code-Kombinationen gebildet und gezählt. Eine Kombination basiert dabei auf dem gemeinsamen Auftreten eines MeSH-Code-Paares in einer Publikation. So bestehen beispielsweise bei zwei Publikationen, in denen die MeSH-Codes [D08.811.913.696.620.682, D12.644.360.450, D12.776.476.450.169.500] für Publikation 1 und [D03.633.100.150.240.190, D03.633.100.150.266.450.206, G04.022, D08.811.913, D12.644.360.450] für Publikation 2 zugeordnet sind, die folgenden Beispiele für Kombinationen und Häufigkeiten [(Kombination aus MeSH-Code, MeSH-Code), Häufigkeit]:

[(D08, D12), 2], [(D03, G04), 1], [(D08, G04), 1], […] auf Ebene 2 und [(D08.811, D12.644),1], [(D08.811, D12.644),2], […] auf Ebene 3.

Zur Visualisierung der folgenden Grafiken wurde Gephi verwendet. Bei der Darstellung wird ein Knoten durch einen MeSH-Code gebildet und die Verbindung beziehungsweise Kante zwischen den Knoten durch die bestehende Kombination aus dem Korpus. Zusätzlich wird die Anzahl des Kombinationsauftretens als

Gewichtung für die Kante verwendet und die Anzahl des absoluten Auftretens eines MeSH-Codes im Korpus als Gewichtung für den Knoten.

Abbildung 36: MeSH-Codes (Ebene 3) mit Auftreten ≥ 5.000



Es ist ersichtlich, dass auf der dritten Ebene vor allem die MeSH-Codes Animals (B01.050), Heterocyclic Compounds, Fused-Ring (D03.633), Heterocyclic Compounds, 1-Ring (D03.383) und Proteins (D12.776) am stärksten vertreten sind und auch die meisten Verbindungen zu anderen MeSH-Codes vorweisen. Besonders hervorgehoben wird das Dreieck zwischen Animals (B01.050), Heterocyclic Compounds, Fused-Ring (D03.633) und Heterocyclic Compounds, 1-Ring (D03.383).

Abbildung 37: MeSH-Codes (Ebene 4) mit Auftreten ≥ 1.800 und Verbindungen ≥ 1.500

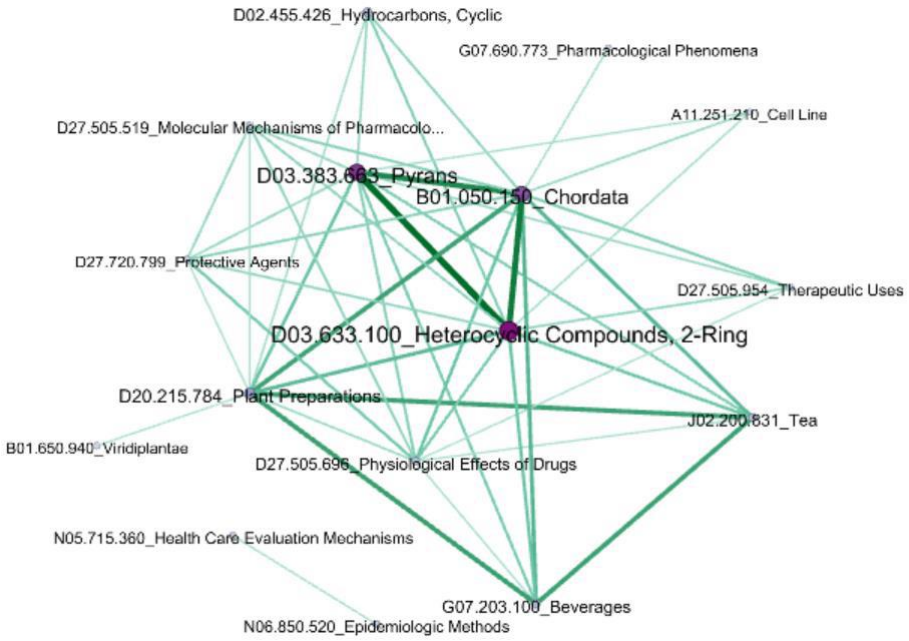
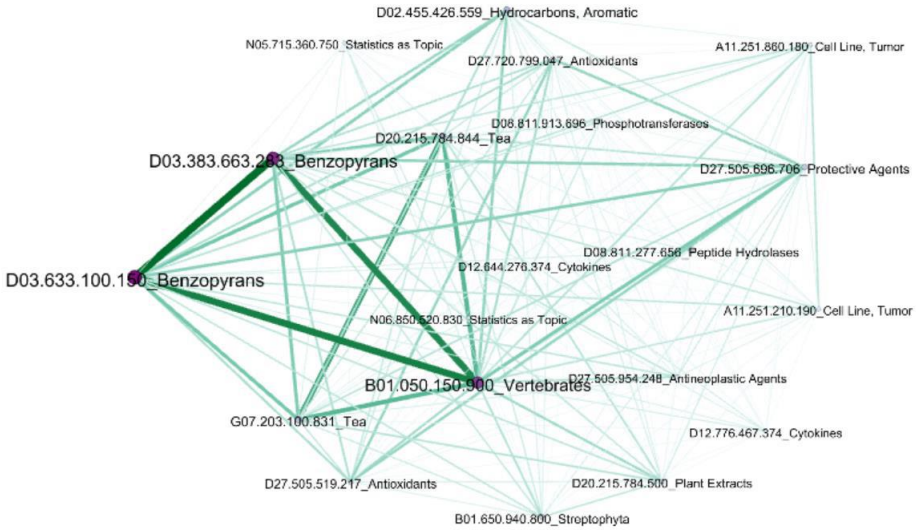


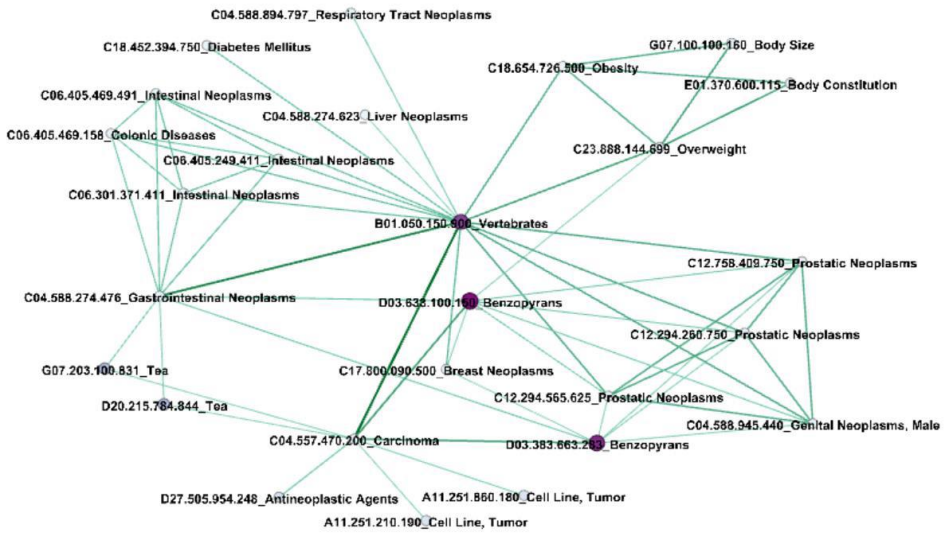
Abbildung 37 ist eingeschränkt auf die MeSH-Codes der vierten Ebene, die mindestens 1.800 Mal aufgetreten sind und mindestens 1.500 Verbindungen zu anderen MeSH-Codes aufweisen. Das Dreieck, welches in der dritten Ebene sichtbar war, wird an dieser Stelle weiter konkretisiert.

Abbildung 38: MeSH-Codes (Ebene 5) mit Auftreten ≥ 1.000



Im Graphen in Abbildung 38 werden MeSH-Codes der fünften Ebene visualisiert, die mindestens 1.000 Mal aufgetreten sind. Die Daten auf dieser Ebene geben erstmalig mehr Aufschluss über mögliche Krankheiten, die in Verbindung stehen, wie A11.251.860.180 und A11.251.210.190 (beide Cell Line, Tumor). Außerdem können die Arten der assoziierten chemischen Substanzen weiter eingeordnet werden.

Abbildung 39: MeSH-Codes (Ebene 5) mit Krankheiten-Kanten und Verbindungen ≥ 150



In Abbildung 39 werden weiterhin die MeSH-Codes der fünften Ebene analysiert, allerdings mit der Einschränkung auf Verbindungen mit Krankheiten, das heißt Hauptgruppe C. Dabei wurden die Verbindungen mit einem Mindestaufkommen von 150 zugelassen. An dieser Stelle wird deutlich, welche Cluster an Krankheiten im Korpus vorhanden sind. Es ist erkennbar, dass ein Großteil der Knoten mit verschiedensten Krebskrankheiten assoziiert ist. Neben den Krebskrankheiten ist auch ein Cluster zu Übergewicht und Fettleibigkeit sichtbar. Auffällig ist weiterhin die starke Verbindung von Vertebrates (B01.050.150.900) beziehungsweise Wirbeltieren zu Carcinoma (C04.557.470.200). Außerdem scheinen Benzopyrans (D03.633.100.150, D03.383.663.283) auch eine bedeutende Rolle zu spielen, da sie als Verbindungspunkt zu verschiedenen Krankheiten stehen.

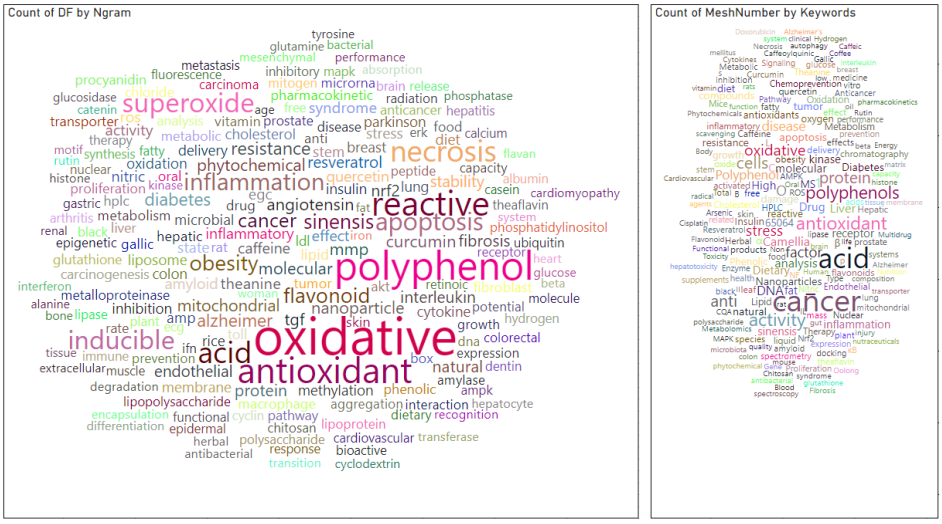
5.3 Wortanalysen

In diesem Abschnitt werden die Abstracts der Artikel und deren Wörter genauer analysiert. Zunächst erfolgt eine Analyse der Wörter mittels N-Grams sowie grundlegende Statistiken. Darauf aufbauend wird eine qualitative Analyse des Korpus mittels Noun Phrases durchgeführt.

5.3.1 N-Grams

In der ersten Analyse der Wörter der vorliegenden Artikel soll mittels N-Grams, das heißt mit Uni- und Bigrams, und NLP-Methoden ein Gefühl für den Inhalt der Publikationen gewonnen werden. Zunächst einmal wurden Unigrams auf Basis des Korpus gebildet und mit Power BI visualisiert. In den folgenden Wordclouds wurden die generierten Unigrams (links) und deren Auftrittshäufigkeit in Artikel-Abstracts visualisiert. Die rechte Wordcloud stellt die (eindeutige) Auftrittshäufigkeit der MeSH-Codes dar, kategorisiert nach Keywords.

Abbildung 40: Wordclouds (Power BI)



Aus der Abbildung ist zu entnehmen, dass Begriffe wie oxidative beziehungsweise antioxidant, polyphenol, reactive, acid und unter anderem cancer am häufigsten in den Abstracts erwähnt wurden und somit die Grundlage für eine tiefere Untersuchung, der Abhängigkeit zwischen diesen Begriffen und grünem Tee, bilden.

Abbildung 41: Wortstatistiken (Power BI)

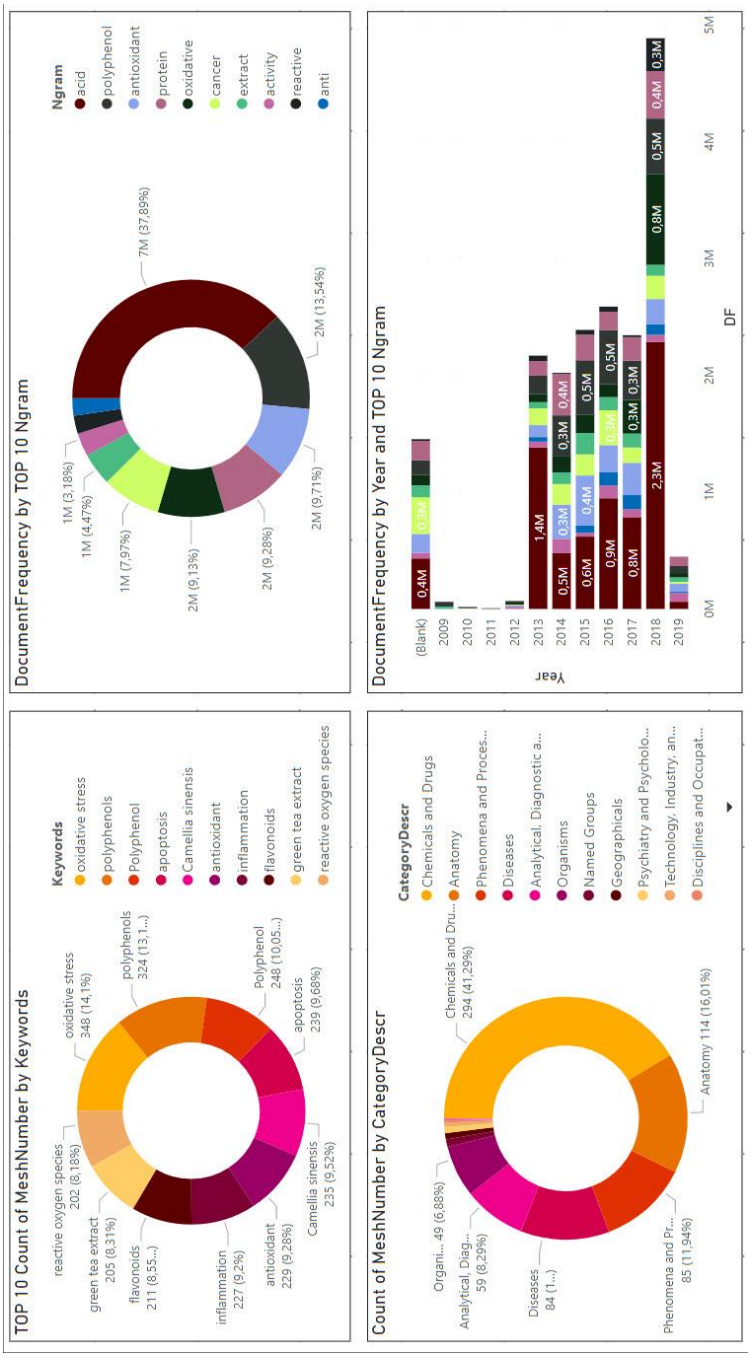
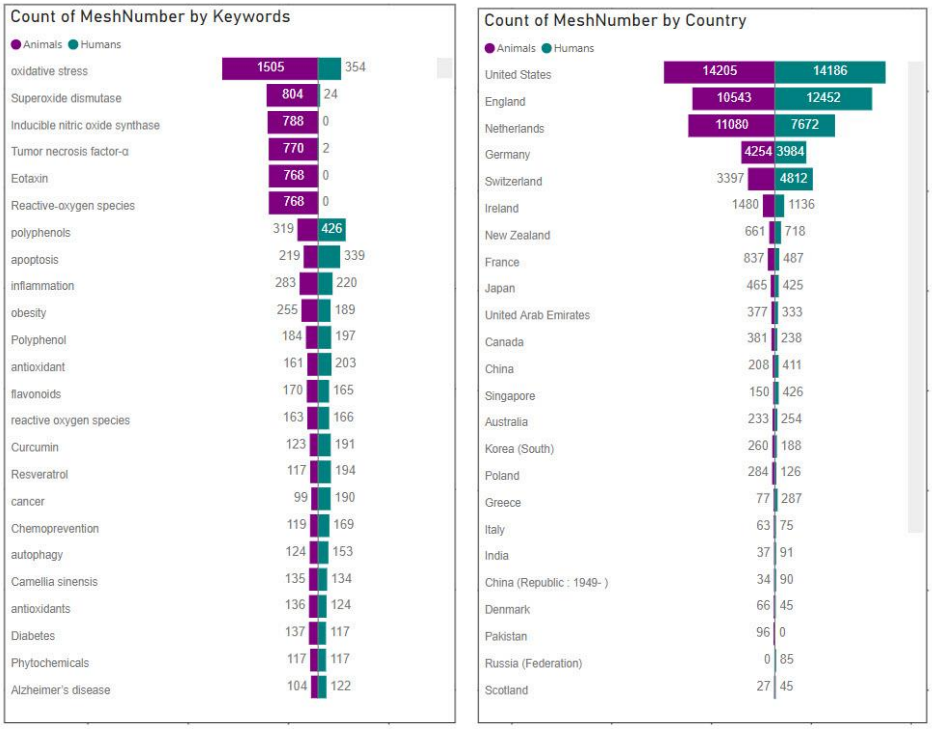


Abbildung 41 veranschaulicht einige allgemeine Metriken bezüglich der Anzahl der verschiedenen Begriffe und deren Verteilung relativ zu den anderen Elementen der jeweiligen Kategorie. Bei den hier abgebildeten Diagrammen wurden die Stoppwörter gefiltert, so dass nur die relevanten Begriffe analysiert werden können. Als erstes können, anhand der zehn am häufigsten auftretenden MeSH-Codes, die Top 10 Keywords ermittelt werden, die am häufigsten in Verbindung mit grünem Tee zu finden sind. Die Top 3 Keywords sind demnach polyphenols, oxidative stress und apoptosis. Wieder anhand der zehn am häufigsten auftretenden MeSH-Codes wurden die MeSH-Hauptgruppen ermittelt, zu denen diese gehören. Die meisten davon fallen in der Kategorie C mit knapp 42 Prozent an gefolgt von Hauptgruppe A und G. Weiterhin, unabhängig von den Kategorien und Keywords, wurden die zehn am häufigsten auftretenden Unigrams ermittelt. Wie bereits erwähnt, wird acid am häufigsten mit grünem Tee assoziiert, gefolgt von polyphenol und antioxidant. Diese Erkenntnis lässt sich auch anhand der abgebildeten Verteilung der Unigrams nach Jahren sehen.

5.3.2 Mensch-Tier-Vergleich

Abbildung 42: Mensch-Tier-Vergleich (Power BI)

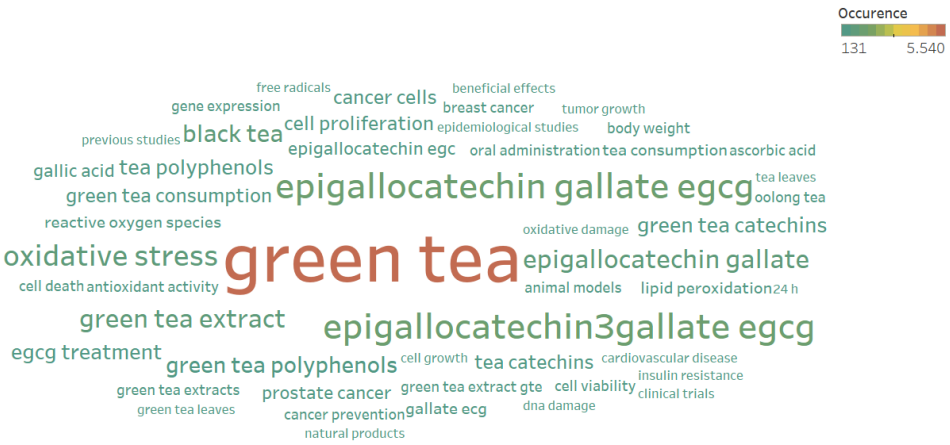


Aus Abbildung 42 lässt sich entnehmen, welche Keywords einen stärkeren Bezug zu Menschen beziehungsweise Tieren haben. Anhand der MeSH-Codes wurde ein Ranking mit den häufigsten Keywords gebildet, sowie deren Verteilung zwischen Mensch und Tier. Dabei wird deutlich, dass der Begriff oxidative stress fast fünf Mal so oft in Bezug zu Tieren als in Bezug zu Menschen erwähnt wird. Zusätzlich treten einige Keywords wie Inducible nitric oxid synthase, Eotaxin und Reactive oxygen species nie in Relation zu Menschen auf, sondern ausschließlich in Bezug zu Tieren. Im Gegensatz dazu sind Begriffe wie inflammation, obesity, antioxidant oder phytochemicals genauso relevant in Bezug auf Menschen wie auf Tiere.

5.3.3 Noun Phrases

Eine weitere Herangehensweise der qualitativen Analyse des Korpus kann mittels Noun Phrases erfolgen. Neben den bereits existierenden menschlichen Annotationen zum Inhalt des Korpus durch Zuordnung von MeSH-Kodierungen, Keywords und Informationen zu den verwendeten chemischen Substanzen, sollen mithilfe von NLP-Techniken, hier Noun Phrases, potenziell weitere Schlüsselbegriffe erfasst werden. Ein Noun Phrase ist eine Substantivgruppe, die syntaktisch zusammengehört, wobei der Kopf dieser Gruppe ein Substantiv sein muss. Der Vorteil von Noun Phrases liegt darin, dass im Vergleich zu einzelnen Tokens und mit der zusätzlichen Berücksichtigung der Syntax potenziell besser und mehr Kontextinformationen erfasst werden können. So wird grüner Tee bei einer reinen Tokenization in grüner und Tee getrennt, aber bei einer Noun Phrase-Extraktion als Ganzes erhalten. Die Datenbasis für die Analyse von Noun Phrases stellt der Korpus mit den beinhalteten Abstract-Texten dar. Die 11.605 Publikationen im Korpus wurden auf 11.306 gefiltert, da einige Publikationen auch leere Abstract-Texte enthalten. Zunächst wird das Auftreten von Noun Phrases im gesamten Korpus analysiert. Als Zweites werden die Noun Phrases publikationsweise analysiert und ein K-Means-Clustering angewendet, um potenzielle Cluster von Publikationen beziehungsweise den dazugehörigen Noun Phrases zu erfassen. Für die Vorbereitung und Analyse wurde Python genutzt. Vor der eigentlichen Extraktion der Noun Phrases werden die Texte vorverarbeitet, so werden diese in Kleinbuchstaben transformiert und spezielle Zeichen, die die Extraktion erschweren könnten (z.B. Klammern), entfernt. Anschließend erfolgt die Extraktion der Noun Phrases mithilfe der Python-Bibliothek spaCy. spaCy bietet eine Standard-Pipeline an, die bereits vordefinierte NLP-Verarbeitungsschritte enthält und auf einen Text angewendet werden kann. Als Beispiel ist die Wortartenerkennung Part-of-Speech-Tagging zu nennen, die maßgeblich für die Extraktion der Noun Phrases ist. Nach der Anwendung der Pipeline entsteht ein Doc-Objekt, welches die Methode noun_chunks enthält, um die erkannten Noun Phrases abzufragen. Zur Visualisierung wurde Tableau verwendet und mehre Wordclouds mit unterschiedlichen Filterungen erzeugt.

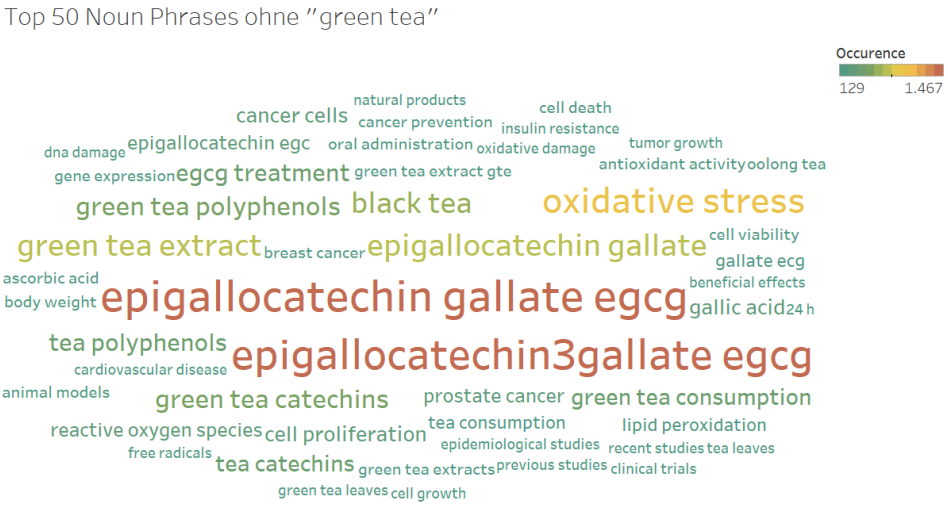
Abbildung 43: Top 50 Noun Phrases



In Abbildung 43 werden die 50 am meisten vorkommenden Noun Phrases visualisiert. Nicht überraschend treten die zu grünem Tee verwandten Begriffe am prominentesten auf, wie zum Beispiel green tea, epigallocatechin gallate egcg, epigallocatechin3gallate egcg und green tea extract. Es wird deutlich, dass bei der Wordcloud mit den meisten Begriffen einige wenige Begriffe mit deutlichem Abstand zu der Mehrheit der restlichen Begriffe auftreten. Die Einschränkung der Top Noun Phrases auf 50 beziehungsweise 30 ändert den Fokus auf zuvor nicht sehr sichtbare Begriffe. So erscheinen nun auch weitere Teesorten wie black tea oder oolong tea und Begriffe zu Krankheiten wie oxidative stress und verschiedene zu Krebskrankheiten wie prostate cancer, cancer cells, tumor growth oder body weight und insulin resistance.

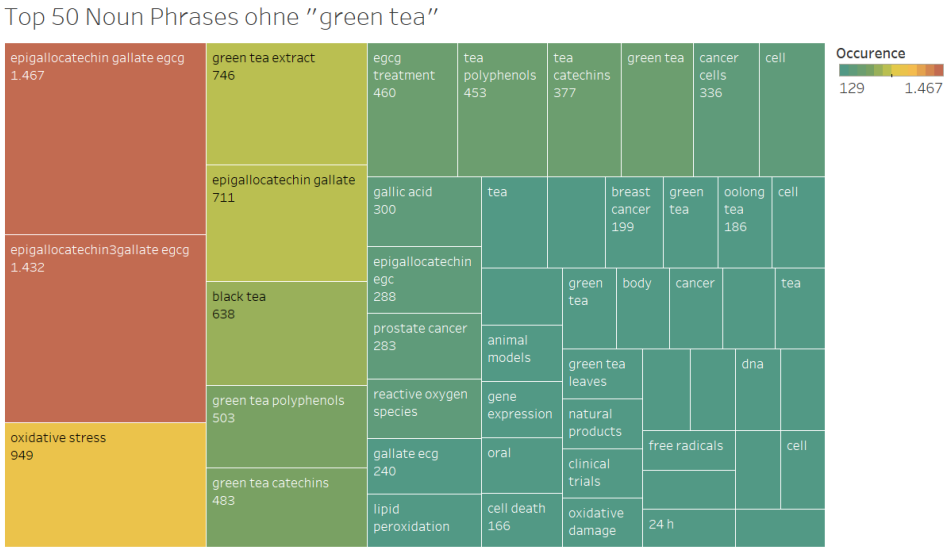
In Abbildung 44 wurde der Begriff green tea ausgeschlossen.

Abbildung 44: Top 50 Noun Phrases ohne „green tea“



In Abbildung 45 wurde ebenfalls der Begriff green tea ausgeschlossen und die Baumkasten-Visualisierung verwendet.

Abbildung 45: Top 50 Noun Phrases ohne „green tea“ (Baumkasten-Struktur)



Noch immer sind Begriffe, die in einem synonymen Verhältnis zum grünen Tee stehen, dominant. Um auch andere Begriffe, die nicht direkt mit grünem Tee oder

dessen prominentesten Inhaltsstoffen verwandt sind, zu analysieren, wurden folgende Begriffe bei der Visualisierung der 50 häufigsten Noun Phrases ausgeschlossen: camellia sinensis, egcg treatment, epigallocatechin egc, epigallocatechin gallate, epigallocatechin gallate egcg, epigallocatechin3gallate egcg, green tea, green tea camellia sinensis, green tea catechins, green tea consumption, green tea extract, green tea extract gte, green tea leaves, green tea polyphenols, tea catechins, tea consumption und tea polyphenols sowie die am häufigsten vorkommende Krankheit in den Noun Phrases, oxidative stress, und weitere Stoppwörter, die für die Analyse uninteressant sind.

Abbildung 46: Top 50 Noun Phrases nach weiterer Begriffsfilterung

Im Folgenden wird der K-Means-Algorithmus angewendet, um mögliche Gruppierungen von Noun Phrases zu erkennen. Auch an dieser Stelle wurde ein Python-Skript geschrieben und dabei die Bibliothek scikit-learn für die Vorbereitung und Anwendung des K-Means-Clusterings verwendet. Nach der Extraktion der Noun Phrases auf Publikationsebene mit spaCy müssen weitere Vorbereitungen für die Anwendung des Algorithmus vorgenommen werden. So müssen die Daten in einer Document-Term-Matrix in vektorisierter Form vorliegen. In unserem Fall stellt ein Document eine Publikation dar und ein Term einen Noun Phrase. Mithilfe des CountVectorizer wird eine Sammlung von Dokumenten in eine Matrix mit dem Auftreten von Terms bzw. Noun Phrases in Dokumenten transformiert. Anschließend kann auf diese Matrix der K-Means-Algorithmus angewendet werden. Der Algorithmus wurde dabei mit jeweils k = 3, 7 und 15 ausgeführt. In den nachfolgenden Visualisierungen sind die Clusterergebnisse mit der Anzahl der

zugeordneten Publikationen zu sehen. Für jedes Cluster wurden außerdem jeweils 20 Noun Phrases angegeben, die das Cluster beschreiben sollen. Dabei sind diese nach der Nähe zum jeweiligen Cluster-Kern sortiert.

Abbildung 47: K-Means-Clustering von Noun Phrases (k=3)

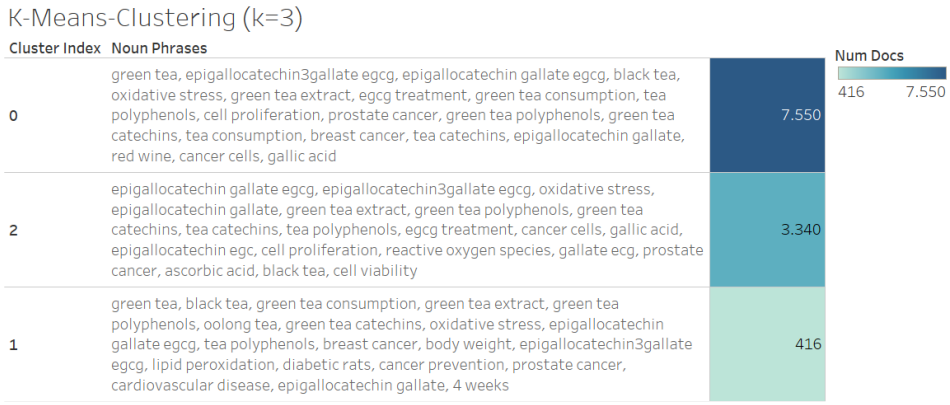


Abbildung 48: K-Means-Clustering von Noun Phrases (k=7)

K-Means-Clustering (k=7)

Cluster Index	Noun Phrases	Num Docs
0	green tea, epigallocatechin3gallate egcg, epigallocatechin gallate egcg, black tea, green tea consumption, egcg treatment, tea polyphenols, oxidative stress, cell proliferation, prostate cancer, green tea extract, green tea polyphenols, tea consumption, green tea catechins, breast cancer, tea catechins, epigallocatechin gallate, red wine, gallic acid, cancer cells	10.078
5	green tea, black tea, green tea consumption, green tea extract, green tea polyphenols, oolong tea, green tea catechins, epigallocatechin gallate egcg, tea polyphenols, breast cancer, body weight, oxidative stress, epigallocatechin3gallate egcg, lipid peroxidation, prostate cancer, cancer prevention, cardiovascular disease, epigallocatechin gallate, diabetic rats, 4 weeks	960
6	epigallocatechin gallate egcg, epigallocatechin gallate, green tea polyphenols, green tea catechins, oxidative stress, tea catechins, tea polyphenols, gallic acid, cancer cells, epigallocatechin egc, gallate ecg, cell proliferation, prostate cancer, egcg treatment, reactive oxygen species, black tea, green tea extract gte, green tea leaves, green tea consumption, tea consumption	124
4	6 months, green tea, prostate cancer pca, 45 assigned green tea drink, prostate specific antigen, daily lycopene n=44, insulinlike growth factor igf peptides, eligible men, unrestricted use, little evidence, 133 men, green tea interventions, 45 assigned placebo, intermediate biomarkers, 15 mg capsules, larger trials.this, green tea supplements, igf binding protein bp2, cancer risk, lycopene supplements	73
2	epigallocatechin3gallate egcg, green tea extract, egcg treatment, oxidative stress, ascorbic acid, cancer cells, green tea catechins, cell proliferation, tea polyphenols, green tea polyphenols, reactive oxygen species, cell viability, cell death, green tea, epigallocatechin egc, gene expression, egcginduced apoptosis, previous studies, normal cells, flow cytometry	69
1	oxidative stress, green tea, epigallocatechin3gallate egcg, egcg treatment, epigallocatechin gallate egcg, reactive oxygen species, lipid peroxidation, green tea polyphenols, free radicals, green tea extract, pc12 cells, cell viability, oxidative damage, diabetic rats, reactive oxygen species ros, antioxidant enzymes, neuronal death, mitochondrial dysfunction, neurodegenerative diseases, cell death	1
3	telomerase activity, pnet cell lines, telomere length, micromolar levels, epigallocatechin gallate egcg, terminal restriction fragment analysis, realtime reverse transcriptasepolymerase chain reaction, green tea, 7 spnet, childhood pnets, telomere lengths, >or= 5fold upregulated htert mrna expression, primitive neuroectodermal tumors, 6 human pnet cell lines, trapnegative d425 cells, normal human cerebellum, 14 normal human brain samples, pnet spnet, telomerase function, htert mrna expression	1

Num Docs
1 10.078

Abbildung 49: K-Means-Clustering von Noun Phrases (k=15)

K-Means-Clustering (k=15)

Cluster Index	Noun Phrases	Num Docs
4	dimethylarginine dimethylaminohydrolase, human umbilical vein endothelial cells, dna methyltransferase, 2 gene, 48 hours, protein level, 2 expression, methylation specific pcr, immunoprecipitationquantitative realtime pcr, human umbilical vein endothelial cells atcc, dna methylation level, methylationsilenced genes, promoter hypermethylation, flow cytometry, dna methyltransferase activity, cancer cells, 2 gene promoter, western blot, chromatin immunoprecipitationquantitative realtime pcr, 2 promoter	9.420
8	eplga membranes, pure plga equivalents, egcg release, nanofiber membranes, approximately 300500 nm, plga degradation, plga membranes, controlled diffusion, controlled drug release, biodegradable nanofiber membranes, activated partial thromboplastin time, 1 week, macroscopic observation, 28 days, green tea, surgical treatment, reduced tissue adhesion, epigallocatechin3ogallate egcg, eplga nanofiber membranes, scavenged reactive oxygen species levels	789
0	collagen membranes, bone regeneration, epigallocatechin3gallate egcg, gbr surgeries, collagen membrane, good biocompatibility, optimal mechanical properties, 12 rats, mechanical properties, surface morphologies, massons trichrome stains, theses membranes, tukeys multiple comparison tests, electron microscope sem, collagen materials, collagen backbone, desirable biological activities, nanohydroxyapatite nanoha, nanoha coatings, nanoha coating	623
6	major metabolites, rat hepatocytes, dog hepatocytes, human hepatocytes, monkey hepatocytes, mouse hepatocytes, glucuronidated m3, green tea catechins, diglucuronidated m5, vitro hepatocytes, certain experimental conditions, chemical oxidation, chemical reactions, glucosidated m2, enzymatic reactions, isomerized m1, methylated m7, methylated m2, isomerized m6, species differences	125
1	green tea, green tea consumption, green tea extract, oxidative stress, green tea polyphenols, epigallocatechin gallate egcg, epigallocatechin3gallate egcg, black tea, breast cancer, lipid peroxidation, green tea catechins, prostate cancer, oolong tea, body weight, tea polyphenols, cancer prevention, cardiovascular disease, diabetic rats, green tea extracts, beneficial effects	107
12	green tea catechins, green tea, dna methylation, oxidative stress, ldl particles, cancer cells, pca risk, lipid peroxides, breast cancer, epigallocatechin3gallate egcg, beneficial effects, epigenetic alterations, human studies, vascular function, 3h]thymidine incorporation, weight loss, green tea extract, epigallocatechin gallate egcg, neointimal formation, green tea intake	89
13	vital mouse brain slices, cultured astrocytes, rna oxidation, oxidative stress, cultured rat astrocytes, ntetraacetic acid, nmethyldaspartic acid nmda receptor activation, cerebral ammonia toxicity, postsynaptic spines, close vicinity, rat brain, acutely ammonialoaded rats, rna colocalizes, splicing protein neurooncological ventral antigen nova2, postsynaptic protein synthesis, cerebral rna oxidation, putative rna transport granules, epigallocatechin gallate, rna species, nicotinamide adenine dinucleotide phosphate	83
5	cell proliferation, green tea, epigallocatechin3gallate egcg, cancer cells, epigallocatechin gallate egcg, tumor growth, mtt assay, oxidative stress, egcg treatment, gene expression, green tea polyphenols, prostate cancer, cell migration, cancer prevention, gelatinase zymography, cell viability, ascorbic acid, epigallocatechin gallate, breast cancer, green tea extract	62

Num Docs: 62 – 9.420

Bei der Darstellung der Cluster-Ergebnisse beim K-Means-Clustering mit k=15 wurden die Cluster mit weniger als 3 zugeordneten Publikationen ausgeschlossen. Beim Analysieren der Cluster-Ergebnisse fällt auf, dass beim Clustering mit k=3 die Cluster anhand der Begriffe nur schwer zu unterscheiden und mit höherem k bessere Abgrenzungen möglich wird. Beispielsweise wird beim Analysieren der Cluster-Ergebnisse bei k=15 deutlich, dass beim Vergleich der fünf ersten

Begriffe der einzelnen Cluster kaum Überschneidungen vorliegen. Trotz des hohen k-Wertes sticht ein bedeutendes Cluster (Cluster 4) mit der Zuordnung zu 9.420 Publikationen stark hervor.

5.4 Krankheiten und Symptome

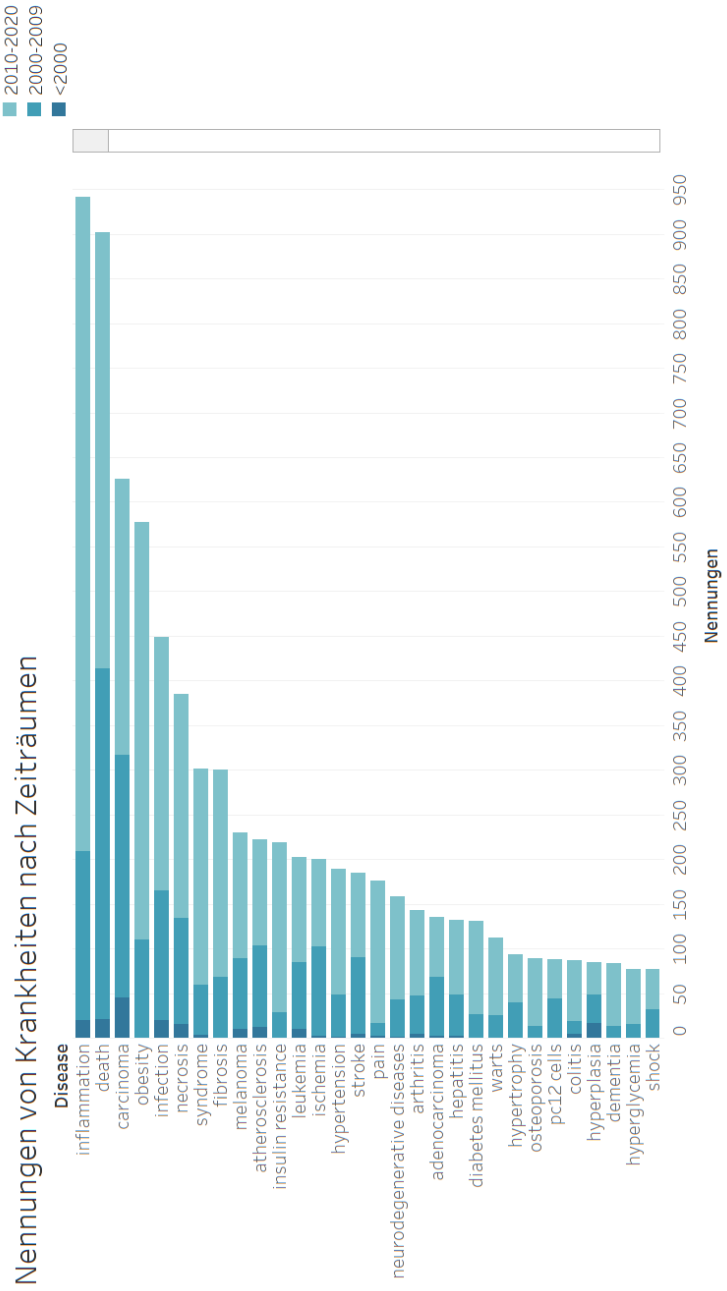
In diesem Abschnitt werden die Krankheiten und Symptome sowie die Publikationsdaten untersucht. Die Datengrundlage stellt die in Kapitel 4.3 (Häufigkeitsanalyse von Krankheiten und Symptomen) durchgeführte Analyse über das Auftreten von Krankheits- beziehungsweise Symptompärchen dar. Dabei wurde das gemeinsame Auftreten von Krankheiten beziehungsweise von Symptomen im Korpus gezählt.

5.4.1 Krankheiten und Symptome im Zeitverlauf

Für die folgende Visualisierung wurden die Trend-Daten der Krankheiten und Symptome verwendet, die in drei Publikationszeiträume unterteilt sind. Die Zeiträume umfassen Daten von (1) vor 2000, (2) zwischen 2000 und 2009 und (3) zwischen 2010 und 2020. Die Daten bestehen immer aus zwei Feldern, einmal die Krankheitsbezeichnung beziehungsweise Symptombezeichnung und einmal die entsprechende Anzahl von Nennungen im Korpus.

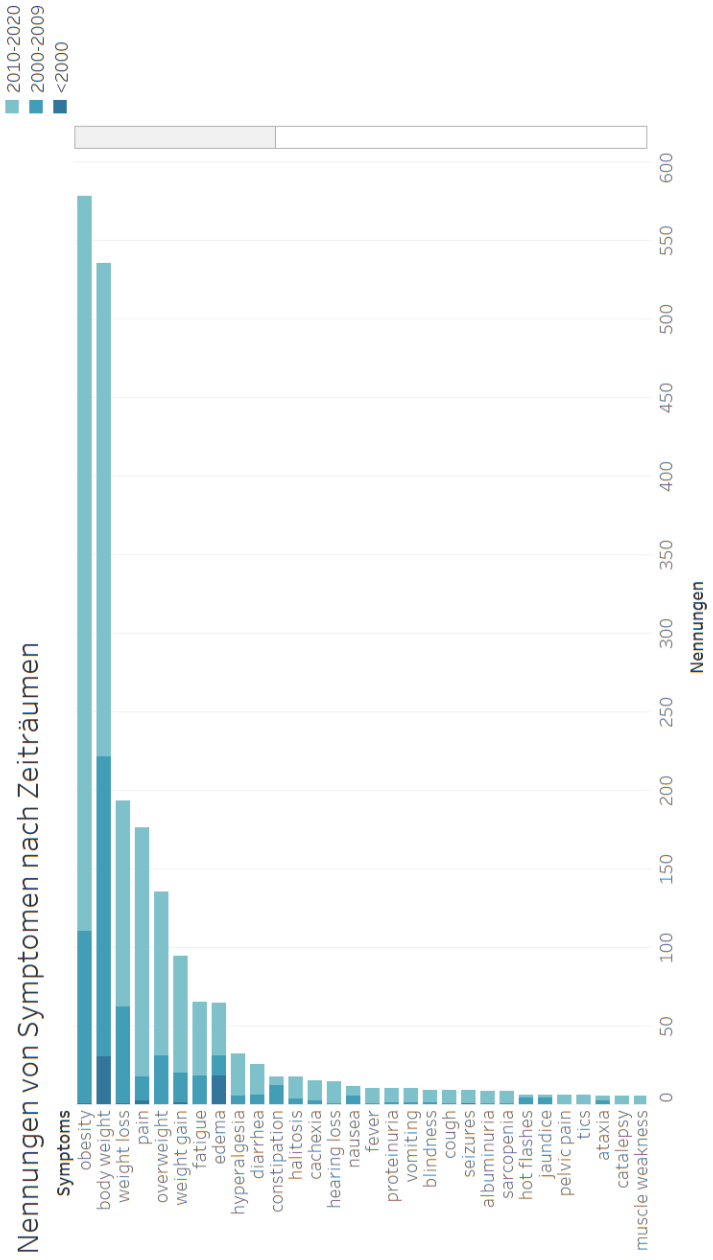
Die Datensets für jeweils die Krankheiten und die Symptome wurden vorher mit Python-Code aneinandergehängt, um eine zusammenhängende Trendanalyse in Tableau zu ermöglichen. Die Visualisierung wurde mit dem Tool Tableau realisiert.

Abbildung 50: Nennungen von Krankheiten nach Zeiträumen



In Abbildung 50 ist ersichtlich, dass nur wenige Publikationen der insgesamt am häufigsten erwähnten Krankheiten vor 2000 (1) vorliegen. Dafür stieg die Anzahl der Publikationen im Zeitraum (3) zum Großteil sehr stark an. Die beiden „Krankheiten“ inflammation (Entzündungen) und death (Tod) mit zwischen 900 und 950 Nennungen werden mit deutlichem Abstand zu den nachfolgenden sehr oft erwähnt. Dabei stieg der Anteil der Publikationen, die inflammation erwähnt haben, im Zeitraum (3) im Vergleich zu den anderen Zeiträumen überproportional an. Ebenso ist eine gleichbleibende Anzahl von Publikationen über death erkennbar. Eine weitere Krankheit, die gleichbleibend in den Zeiträumen (2) und (3) präsent ist, ist die Krankheit carcinoma (Krebs).

Abbildung 51: Nennungen von Symptomen nach Zeiträumen



Schaut man sich das Balkendiagramm zu den Symptomen an, fallen die „Symptome“ obesity (Fettleibigkeit) und body weight (Übergewicht) auf, da diese ebenfalls mit einem sehr deutlichen Abstand zu den anderen Symptomen erwähnt werden. Das Symptom obesity wird im Zeitraum (3) fast vierfach so oft genannt wie in (2). Außerdem scheint das „Symptom“ body weight auch vor 2000 ein wichtiges Thema gewesen zu sein. Bei einer Mindestpublikationsanzahl von 100 gehören die „Symptome“ obesity, body weight, weight loss, pain und overweight zu den am häufigsten genannten Symptomen.

Für die Vorbereitung der Daten für die Gephi-Visualisierung wurden Python und die Bibliotheken Pandas und insbesondere Networkx verwendet. Um die Daten für die Graphen-Visualisierung nicht manuell erzeugen zu müssen, wird mithilfe von Networkx eine GML-Datei erzeugt, die von Gephi gelesen werden kann. Dafür werden zunächst die Rohdaten in Form von eindeutigen Knoten und Kanten transformiert. In unserem Fall stellen die Krankheiten bzw. Symptome die Knoten und die Pärchen die Kanten dar. Die Anzahl der Nennungen der Pärchen wird für die Visualisierung auf Werte zwischen 1 und 15 skaliert und als Gewicht für die Kanten, also der Beziehung zwischen zwei Krankheiten bzw. Symptomen genutzt. Mithilfe der Knoten, Kanten, und Kantengewichtungen werden in Networkx ein Graph-Objekt aufgebaut, welches anschließend als GML-Datei exportiert wird.

Für die Krankheits- und Symptompärchen wurden jeweils eine Graphen-Visualisierung in Gephi erstellt und anschließend aufgrund der Größe der Graphen auf Teilgraphen gefiltert. Die Knotengröße, Knotenfarbe und Knotentextgröße ist dynamisch nach dem Grad bzw. Degree des Knotens definiert, wobei der Grad die Anzahl der Verbindungen eines Knotens zu anderen Knoten darstellt. Bei den Kanten werden die Kantendicke und Kantenfarbe nach dem Kantengewicht angepasst.

Beim Vergleich der beiden Gesamtgraphen Abbildung 52 und Abbildung 53 wird deutlich, dass der Graph über Krankheiten viel dichter ist und ein relativ starkes Zentrum aufweist. Um den Graphen einzugrenzen, wird der Graph nach Knoten gefiltert, die mindestens einen Grad von 50 haben, also mindestens 50 Verbindungen zu anderen Knoten haben.

Abbildung 52: Gesamtgraph über Krankheiten

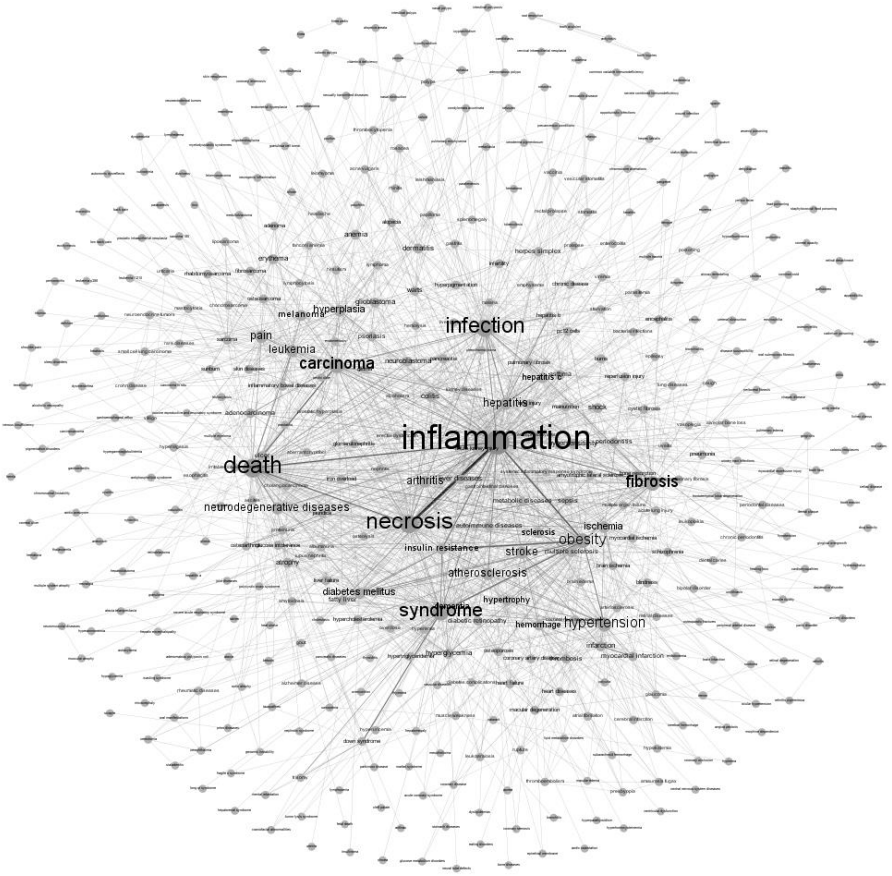


Abbildung 53: Gesamtgraph über Symptome

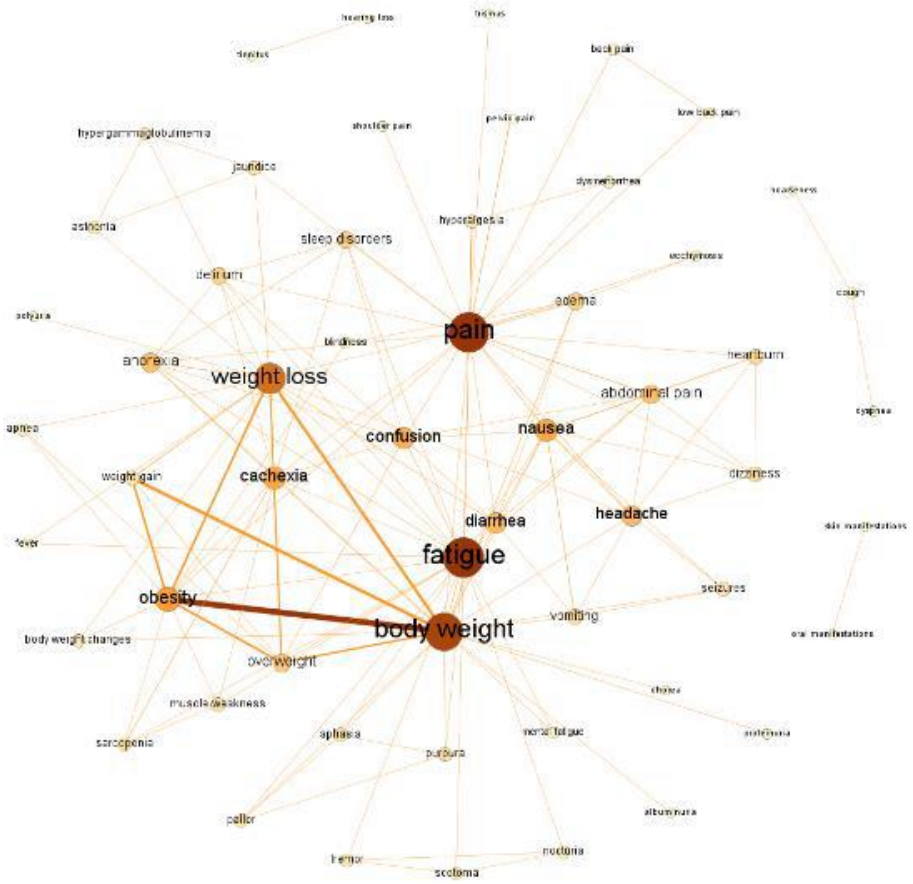
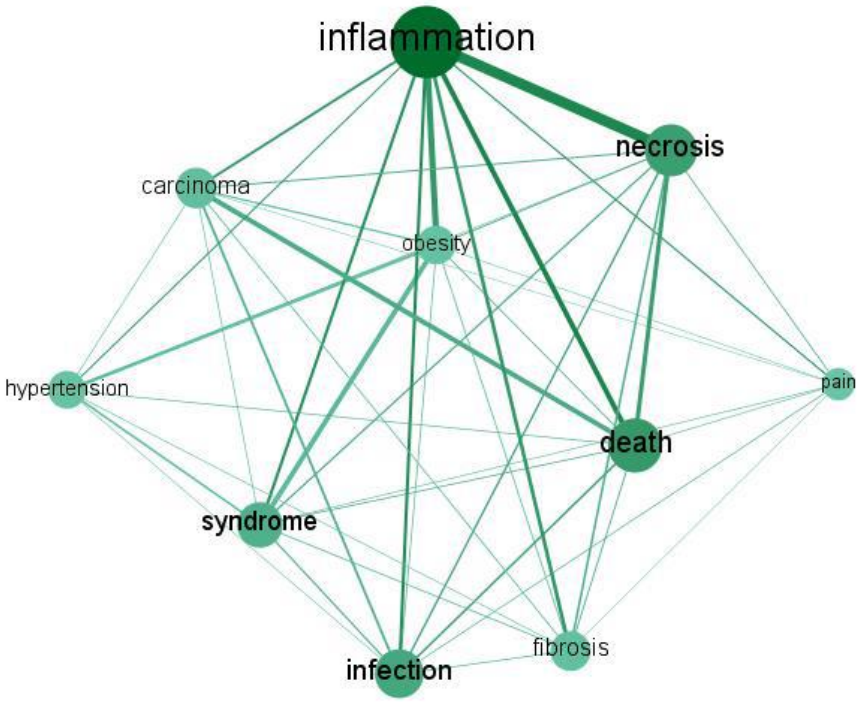


Abbildung 54: Teilgraph über Krankheiten

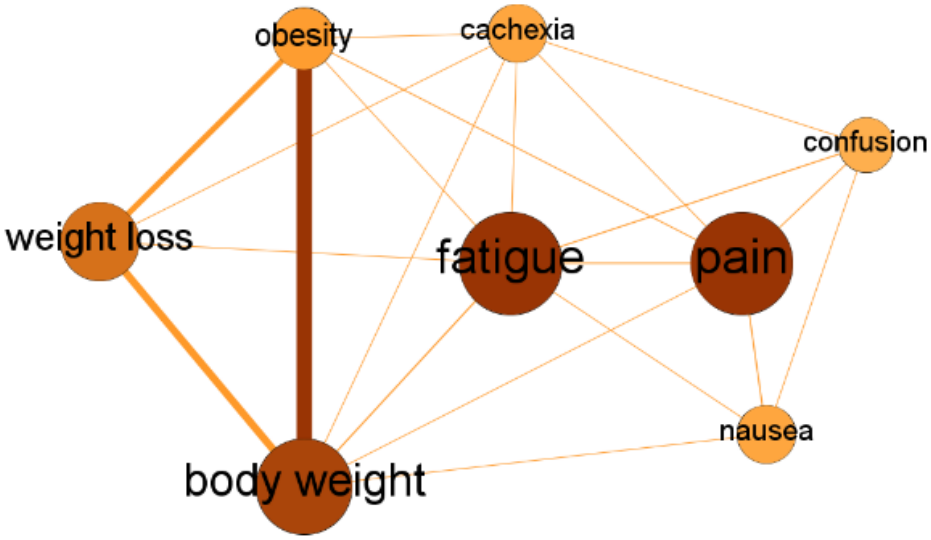


Das Ergebnis ist deutlich lesbarer und zeigt, dass die Krankheiten inflammation und necrosis (Nekrose) im Korpus im Zusammenhang mit grünem Tee oft zusammenhängend betrachtet werden. Interessant ist auch die Verbindung zwischen carcinoma und death sowie die relativ starken Verbindungen von obesity zu inflammation und syndrome.

Im Gesamtgraphen der Symptomvorkommen ist zunächst offensichtlich aufgrund der Stärke der Verbindungen und des Grades ein klares Cluster der „Symptome“ body weight, weight loss, obesity und weight gain zu erkennen. Die „Symptome“ pain, fatigue und body weight haben alle gemeinsam, dass sie einen hohen Grad haben, das heißt die höchste Anzahl an Verbindungen zu anderen Knoten. Die Symptome pain und fatigue haben interessanterweise einen hohen Verbindungscharakter, aber keine starke Verbindung zu bestimmen Symptomen.

Der folgende Teilgraph für die Krankheiten besteht aus Knoten, die einen Grad von mindestens 10 (von max. 24) haben.

Abbildung 55: Teilgraph über Symptome



5.4.2 Publikationsdaten

Die folgende Analyse zielt darauf ab, Autorinnen und Autoren, Publikationen, Journals und Länder zu vergleichen und das qualitative Aufkommen der Publikationen einem qualitativen Kriterium (H-Index) gegenüberzustellen. Sowohl die Aufbereitung als auch die Visualisierung der Daten erfolgt mit Tableau Desktop.

Die verwendeten Datenquellen werden überwiegend als CSV-Dateien bezogen und beinhalten Metadaten zu den Publikationen des zugrundliegenden Korpus. Darin sind je nach Datei jahresbezogene Informationen über die Anzahl an Publikationen pro Autor, pro Journal und pro Land aufgeführt. Desweiteren beinhalten diese auch Angaben zu den Publikationen zugeordneten MeSH-Codes. Zur Datenvorverarbeitung werden die CSV-Dateien teilweise in das MS-Excel-Format (.xlsx) umgewandelt. Dies ist beispielsweise zur korrekten Dimensionierung der H-Index-Angaben des Journals erforderlich, da sie in der Ausgangsdatei um den Faktor 10 erweitert sind. Weitere Schritte der Datenvorverarbeitung können direkt in Tableau Desktop durchgeführt werden. Diese beinhalten die Datentyp-Festlegung (z.B. Jahreszahlen als Datum), die Umwandlung von Dimensionen in Kennzahlen (z.B. Pmid [PubMed-ID]) und die situationsabhängige Kennzahl-Wertdefinition (Summe, Mittelwert, Median, usw.).

Zunächst werden die statistischen Daten zu den Autorinnen und Autoren, Publikationen, Journals und Ländern verglichen. Anschließend werden die Daten kumuliert analysiert und durch weitere Faktoren wie H-Index und MeSH-Code differenziert. Mit Hilfe dieser Analysen können die Publikationspotenzen der einzelnen Länder, Journals und Autorinnen und Autoren ermittelt und bewertet werden. Um länderspezifische Einflüsse betrachten zu können, werden im Anschluss die Publikationsverläufe der Länder mit deren durchschnittlichem H-Index Verlauf verglichen.

5.4.2.1 Statistik der Publikationsdaten

Als erstes werden die Verläufe der Anzahl an publizierenden Ländern, Journals, Publikationen, und Autoren visualisiert. Dazu werden diese Werte sowohl als Linien- als auch als Bereichsdiagramme dargestellt, wodurch die Verläufe der einzelnen Kennzahlen in Bezug auf deren Wachstum bewertet werden können.

Abbildung 56: Verlauf der publizierenden Länder, Journals, Publikationen und Autoren

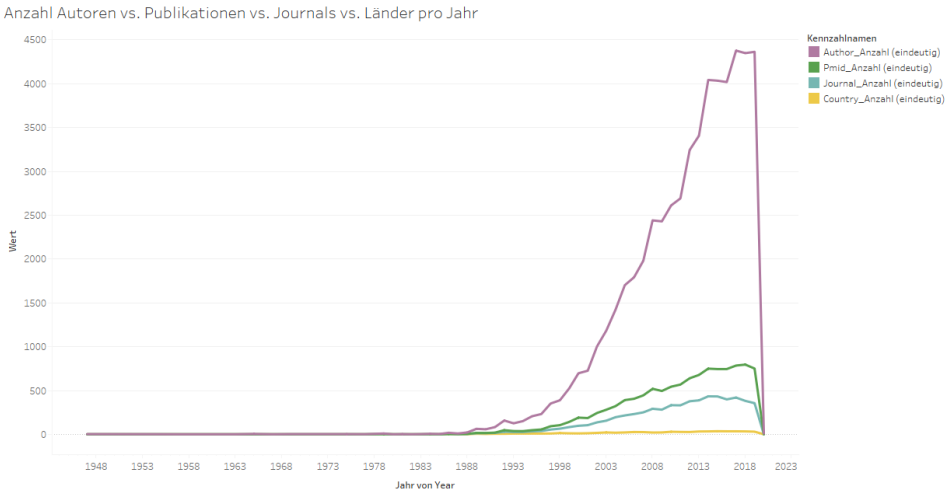


Abbildung 56 zeigt die Verläufe der Anzahl an publizierenden Ländern (unterster Graph), Journals (zweiter Graph), Gesamtpublikationen (dritter Graph) und Autoren (oberster Graph) als Liniendiagramm an. Auf der x-Achse wird das Zeitintervall von 1948 bis 2020 und auf der y-Achse die Anzahl der Kennzahlen in 500er-Schritten abgebildet.

Abbildung 57: Verlauf der publizierenden Länder, Journals, Publikationen und Autoren (einzeln)

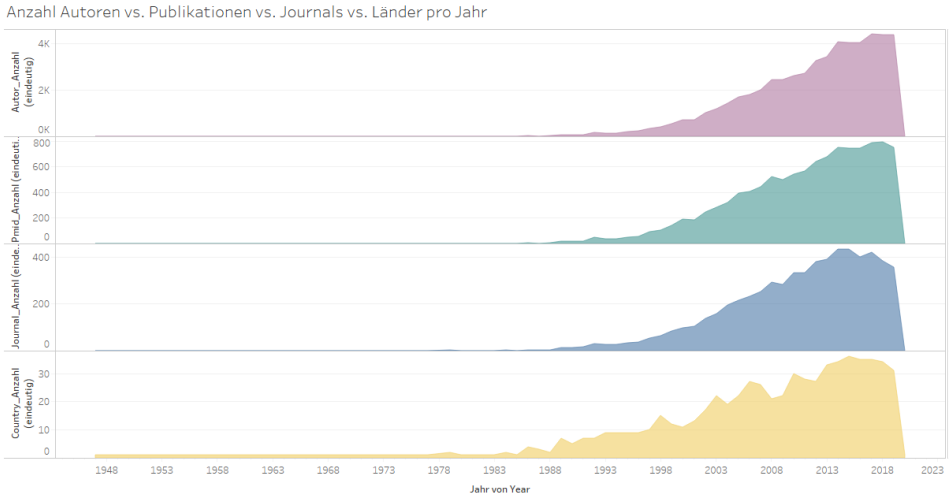


Abbildung 57 zeigt die gleichen Kennzahlen und deren Ausprägungsverläufe als separierte Flächendiagramme. Von oben nach unten betrachtet sind die Verläufe der Anzahl von publizierenden Autorinnen und Autoren, Publikationen, Journals und Ländern dargestellt.

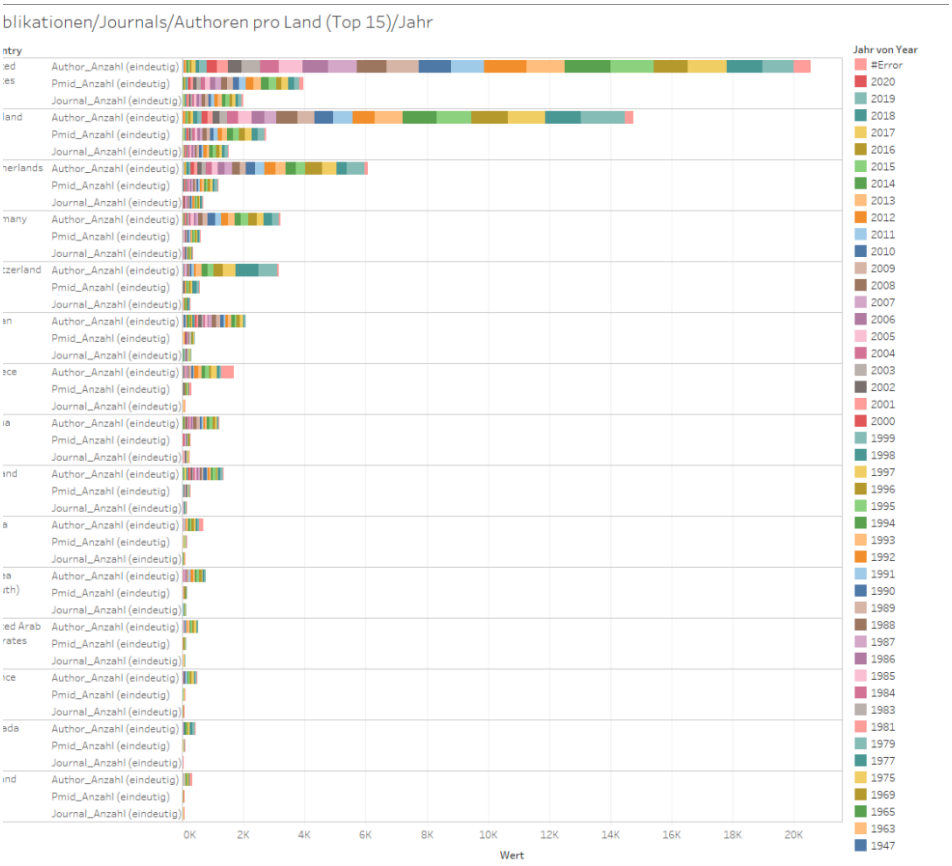
5.4.2.2 Publikationspotenz

Für die Analyse in diesem Abschnitt wird die Publikationspotenz durch die Anzahl an unterschiedlichen Autorinnen und Autoren, Publikationen und Journals bestimmt. Dabei wird zuerst die Gesamtzahl der Publikationen pro Land in einer Heatmap dargestellt. Zur genaueren Analyse der Publikationspotenzen werden anschließend die Kennzahlenverläufe der einzelnen Länder als Balken- und im weiteren Verlauf zusammen mit dem durchschnittlichen H-Index auch als Liniendiagramm dargestellt. Danach werden Potenzanalysen für die Autorinnen und Autoren und Journals durchgeführt. Jedoch wird dabei im Gegensatz zur Potenzanalyse der Länder nur noch die Anzahl an Publikationen und deren Verlauf pro Autor bzw. Journal betrachtet. Dazu werden, wie bei der Länderanalyse, Balken- und Liniendiagramme für jeden Autor und jedes Journal erstellt.

Abbildung 58: Heatmap der Publikationszahlen aller im Korpus enthaltenen Länder

Abbildung 58 zeigt die Heatmap zu den Gesamtpublikationszahlen aller im Korpus enthaltenen Länder. Die Farbe der Länder symbolisiert die jeweilige Anzahl an Publikationen zum Thema grünem Tee. Dabei reicht das Spektrum der Farbskala von 1 (grün) bis 3.951 (rot).

Abbildung 59: Anzahl der Autoren, Publikationen und Journals der Top 15 Länder pro Jahr



Werden die Kennzahlen der Top 15 publizierenden Länder pro Jahr als Balkendiagramm dargestellt, ergibt sich Abbildung 59: Anzahl der Autoren, Publikationen und Journals der Top 15 pro Jahr. Die Farbskala zeigt die Zuweisung der jeweiligen Jahreszahl zu einer Farbe. Dabei ist zu beachten, dass der Wert #Error der Farbskala stellvertretend für Kennzahlwerte der Länder ohne Jahreszahleintrag steht. Die Länder sind nach der Anzahl ihrer Gesamtpublikationen absteigend sortiert. Zu jedem der Länder sind je drei Balkendiagramme dargestellt. Von oben nach unten zeigen diese die Ausprägungen der Autorinnen und Autoren, Publikationen und Journal-Anzahl des jeweiligen Landes.

Abbildung 60: Anzahl der Publikationen der Top 30 Autoren pro Jahr

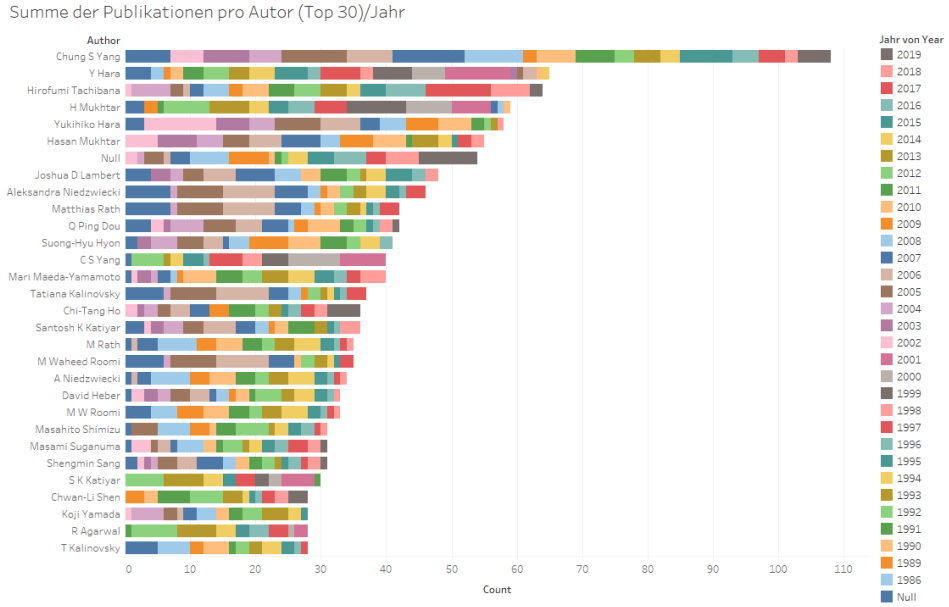


Abbildung 60 zeigt die Anzahl der Publikationen der Top 30 Autorinnen und Autoren pro Jahr als Balkendiagramm. Mit Hilfe der Farbskala lässt sich die Verteilung der Publikationen auf die unterschiedlichen Jahre nachvollziehen. Die Autoren sind absteigend nach der Gesamtzahl ihrer Publikationen sortiert.

Zur Ergebnisvalidierung wurden die Verläufe der Publikationen der Top 15 Autoren als Liniendiagramme dargestellt. Dadurch ist erkennbar, dass einige Autoren, aufgrund einer unterschiedlichen Schreibweise ihrer Namen, separat aufgeführt werden. Beispielsweise publizierte C. S. Yang von 1992 bis 2002 und Chung S. Yang von 2002 bis 2018. Diese Fälle müssen bei weiteren Analysen zu Autorinnen und Autoren zusammengefasst werden.

Abbildung 61: Anzahl der Publikationen der Top 30 Journals pro Jahr

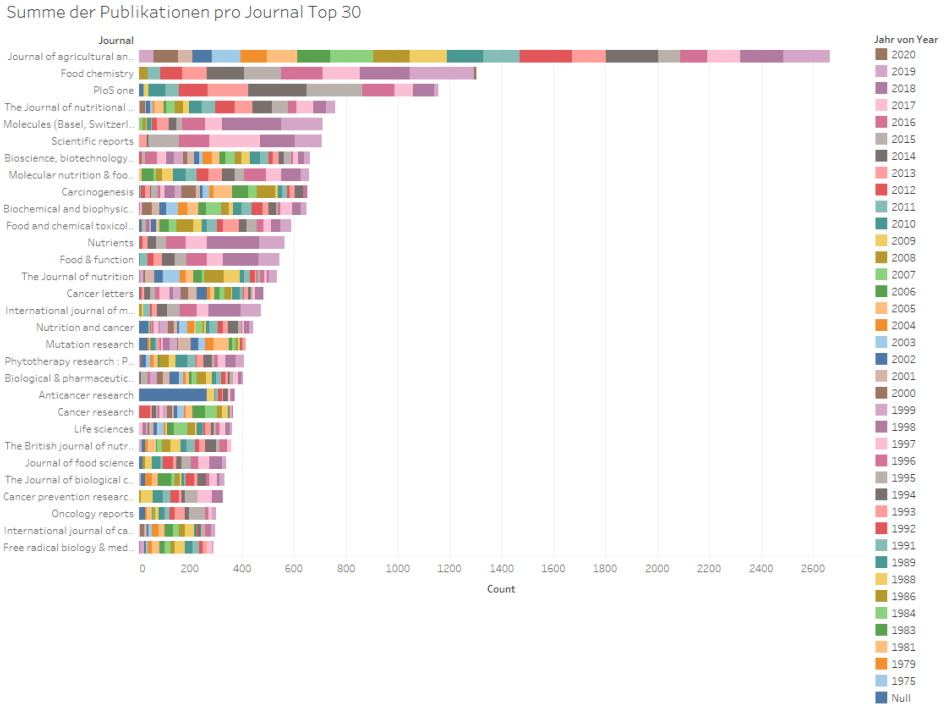


Abbildung 61 zeigt die Anzahl der Publikationen der Top 30 Journals pro Jahr als Balkendiagramm. Mit Hilfe der Farbskala lässt sich die Verteilung der Publikationen auf die unterschiedlichen Jahre nachvollziehen. Die Journals sind absteigend nach der Gesamtzahl ihrer Publikationen sortiert.

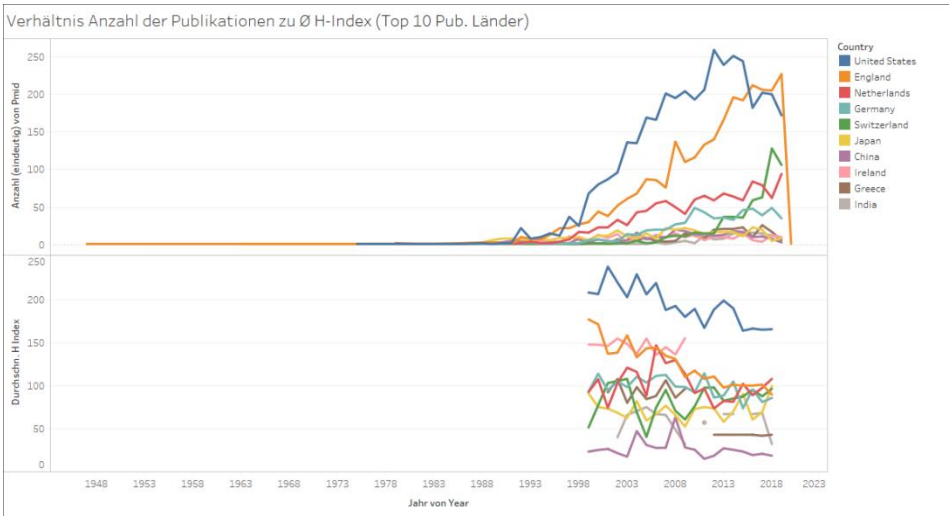
5.4.2.3 H-Index

Um die wissenschaftliche Aussagefähigkeit und den Publikationsverlauf des Green Tea-Korpus bewerten zu können, werden der H-Index der Journale und der allgemeine Zuwachs an Publikationen in der PubMed-Datenbank als Vergleichsgrößen in die Analyse miteinbezogen. Dazu werden zuerst Liniendiagramme mit den Verläufen zu den Publikationen von PubMed und denen des Korpus, sowie zum durchschnittlichem H-Index der Korpus-Publikationen pro Jahr erstellt. Durch diese Diagramme lassen sich individuelle Besonderheiten im Verlauf der einzelnen Kennzahlen vom Allgemeinen differenzieren.

Der H-Index gilt als ein Maß für die wissenschaftliche Tragweite eines Autors oder einer Autorin bzw. Journals. Braun, Glänzel und Schubert definieren den Hirsch-type index for journals wie folgt (vgl. Braun et al., 2006):

> *„We suggest that a h-type index would advantageously supplement journal impact factors at least in two aspects: (i) it is robust, i. e., insensitive to an accidental excess of uncited papers and also to one or several outstandingly highly cited papers; (ii) it combines the effect of „quantity" (number of publications) and „quality" (citation rate) in a rather specific, balanced way (thereby, it is expected to reduce the apparent „overrating" of some small review journals). The journal h-index would not be calculated for a „life-time contribution" (as suggested by Hirsch for individual scientists), but for a definite period – in the simplest case for a single year."*

Abbildung 62: Verlauf der Anzahl an Publikationen und Ø-H-Index der Publikationen pro Land



Bildet man den Verlauf der Publikationen und den Ø des H-Index der Journale, in denen publiziert wurde, der Top 10 Länder als Liniendiagramm ab, so ergibt sich Abbildung 62. Das Diagramm zeigt den Verlauf der MEDLINE-Zitationen im Vergleich zu den Publikationen des Korpus und deren Ø-H-Index (gelb).

Den Abschluss der Analysen bildet die Auswertung der MeSH-Codes. Dabei werden die Codes sowohl auf ihre Häufigkeit als auch ihre länderspezifischen Ausprägungen und ihren Einfluss auf den H-Index analysiert. Für die Analyse der

MeSH-Codes werden additive Balkendiagramme und Bubblecharts verwendet. Als Exkurs wird noch ein Balkendiagramm mit der Verteilung der MeSH-Codes mit den höchsten Ø-H-Indizes und mindestens zehn Gesamtpublikationen der Top 10 Publikationsländer erstellt. Somit lässt sich eine Bestenauslese der MeSH-Codes erstellen und deren länderspezifischer Einsatz ermitteln.

Abbildung 63: Summe der Publikationen, Ø-H-Index und Verteilung pro Top 10 Länder der Top 40 (Pub.) MeSH-Codes

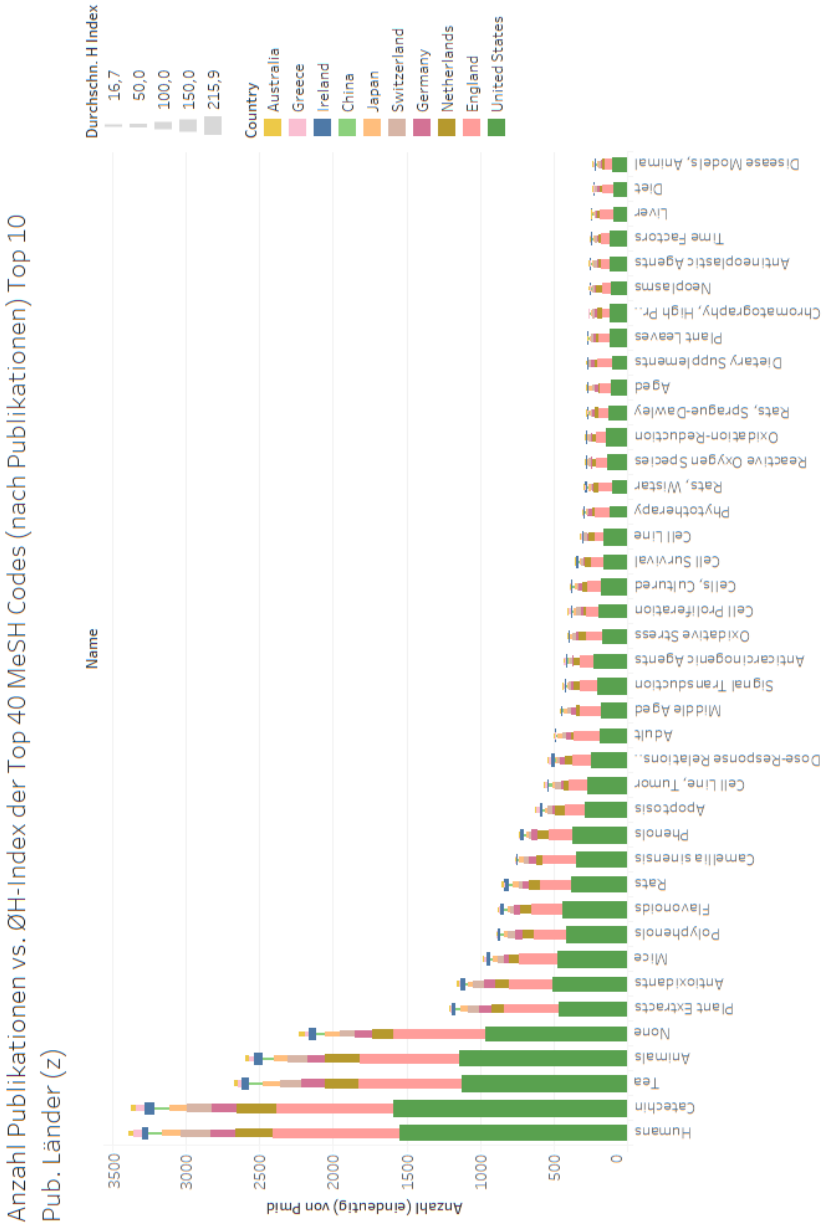
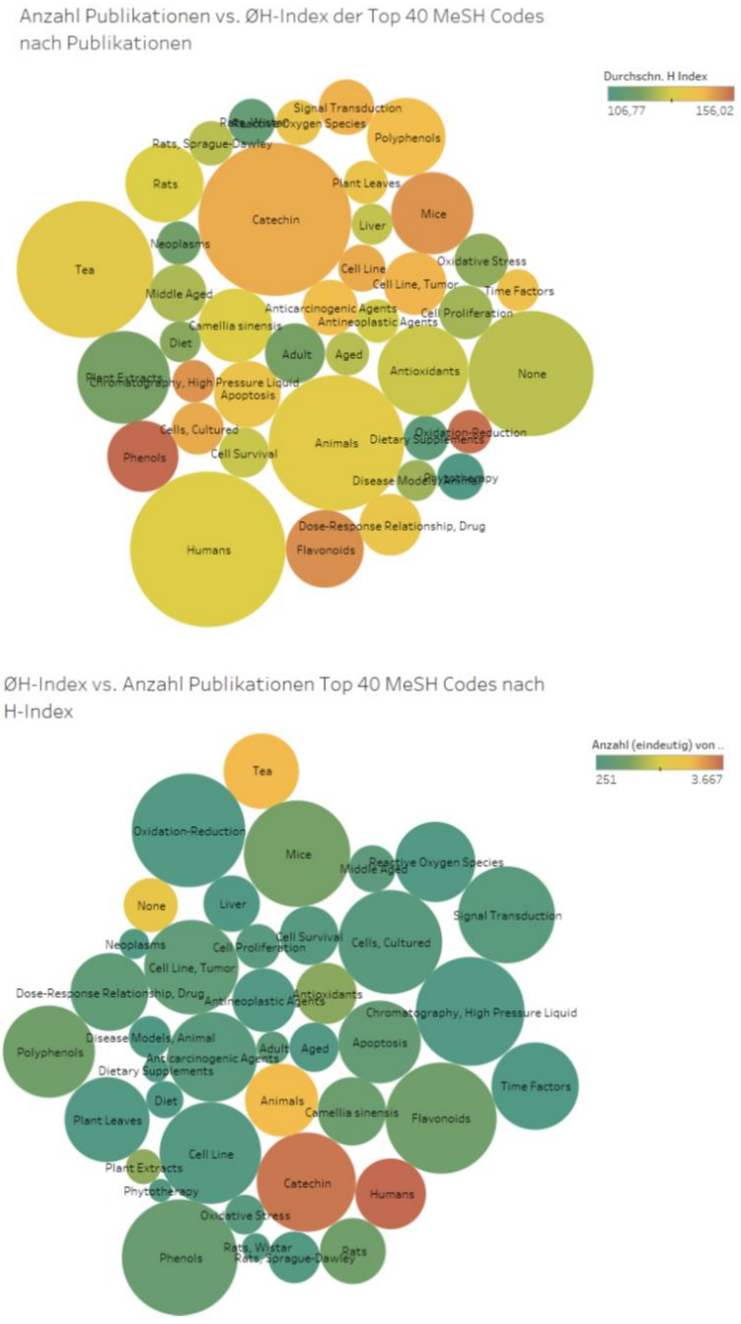


Abbildung 63 zeigt die Anzahl an Publikationen (y-Achse) und den zugehörigen Ø-H-Index der publizierenden Journals (Balkendicke) für die Top 40 publizierten MeSH-Codes (x-Achse) der Top 10 Publikationsländer (Farbe) als absteigend sortiertes Balkendiagramm.

Abbildung 64: Anzahl an Publikationen und Ø-H-Index der Top 40 Pub. MeSH-Codes



Die Anzahl der Publikationen lässt sich anhand des Top-MeSH-Codes sowie des entsprechenden H-Indexes im Bubblechart (ohne Länder) betrachten. Die Dicke der Bubbles repräsentiert die Anzahl der Publikationen und die Farbe den entsprechenden Ø-H-Index. Dabei geht die Skala von 106,77 (Grün) bis 156,02 (Rot) (oberes Diagramm).

Andersherum betrachtet repräsentiert die Dicke der Bubbles den Ø-H-Index und die Farbe die entsprechende Anzahl der Publikationen. Dabei geht die Skala von 251 (Grün) bis 3.667 (Rot). (unteres Diagramm).

Abbildung 65: Ø-H-Index, Summe der Publikationen und Verteilung pro Top 10 Länder der Top 40 (Pub.) MeSH-Codes

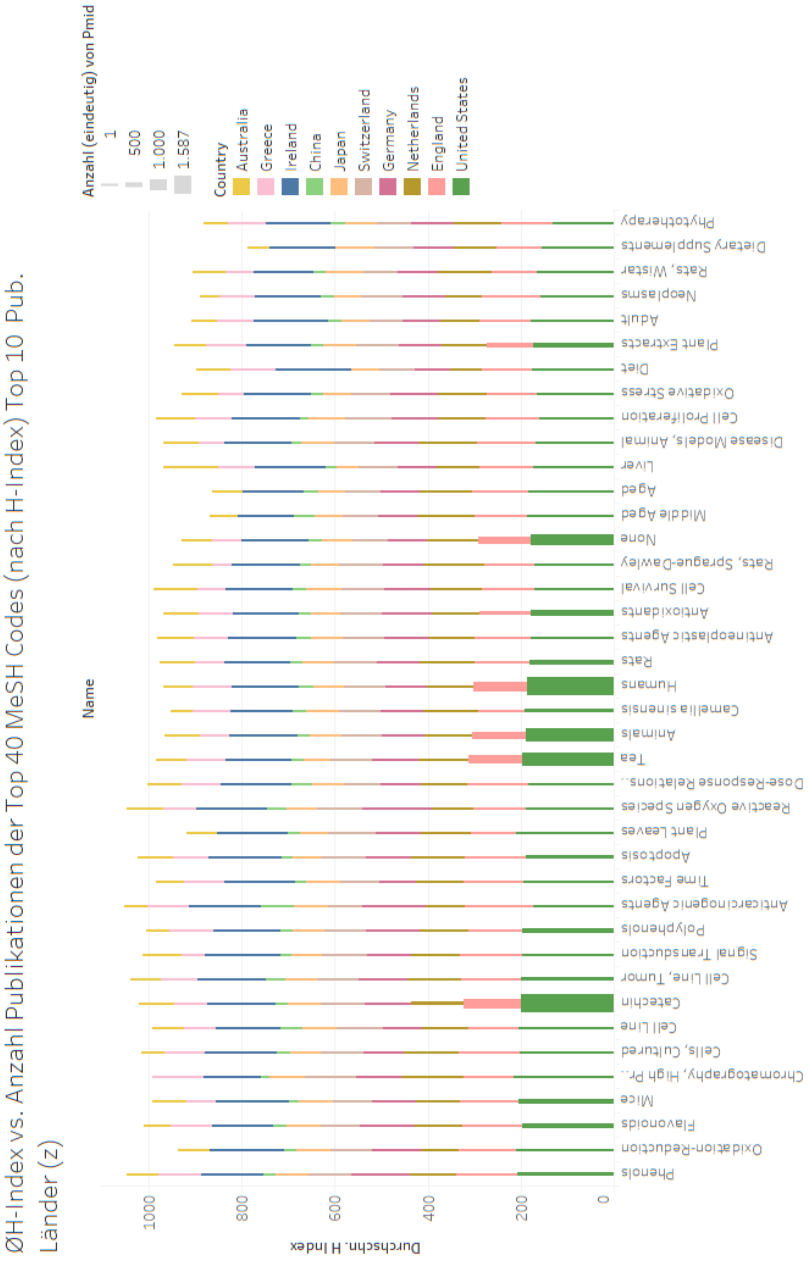


Abbildung 65 zeigt den Ø-H-Index der publizierenden Journals (y-Achse) und die zugehörige Anzahl an Publikationen (Balkendicke) für die Top 40 H-Index MeSH-Codes (x-Achse) der Top 10 Publikationsländer (Farbe) als absteigend sortiertes Balkendiagramm.

5.5 Vergleich des Datenkorpus mit der gesamten PubMed-Datenbank

Die folgenden Darlegungen zeigen, inwiefern sich die wissenschaftliche Behandlung des Themas grüner Tee in der Medizin von der Gesamtheit der PubMed-Datenbank ohne thematischen Schwerpunkt abhebt. Insbesondere soll überprüft werden, ob bestimmte Krankheiten oder Symptome verstärkt mit grünem Tee behandelt werden und ob es somit „typische“ medizinische Einsatzgebiete für grünen Tee gibt. Sollte sich dies im weiteren Verlauf der Untersuchung nicht feststellen lassen und sich die Verteilungen mit und ohne Betrachtung des grünen Tees decken, kann dies ein Indiz für Arbitrarität sein.

Als Grundlage dienen repräsentativ für die PubMed-Datenbank im Allgemeinen die Anzahlen der jeweiligen Krankheiten und Symptome, wie sie im Nature-Artikel von Zhou et al. (2014) „Human symptoms-disease network“ erfasst wurden. Dazu gehören die Dateien diseases.csv und symptom.csv respektive. Als Daten im Bezug zum grünen Tee wurde aus den Abstracts des Datenkorpus eine textbasierte Extraktion von Krankheiten und Symptomen durchgeführt. Dazu wurden die Abstracts nach Bezeichnungen von MeSH-Codes durchsucht und die entsprechenden Bezeichnungen jeweils gezählt. Der Vergleich zu den MeSH-Codes, mit denen die Artikel zum grünen Tee in der PubMed-Datenbank versehen wurden, führte hierbei zu keinen aussagekräftigen Ergebnissen, wie später dargelegt wird.

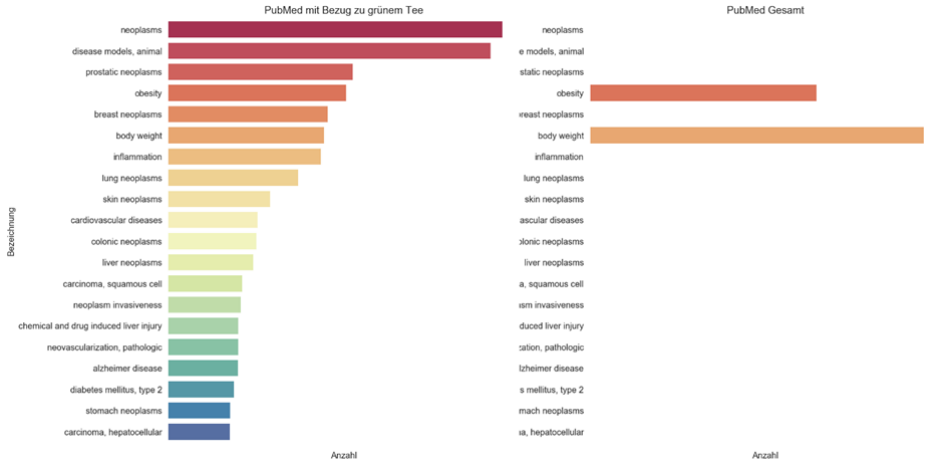
Zur Aufbereitung der Daten wird Python 3.7 mit der Bibliothek pandas verwendet. Darin werden die Ursprungstabellen zu Krankheiten und Symptomen sowie MeSH-Codes im Bezug zu grünem Tee jeweils mit ihrem gesamtheitlichen Gegenstück per Join verbunden. Im Falle der MeSH-Code-Tabelle werden hierbei nur die MeSH-Codes des Typs C berücksichtigt. Als Resultat entstehen darin jeweils Tabellen, welche die wörtliche Bezeichnung eines MeSH-Codes und die dazugehörigen absoluten Zählungen jeweils mit und ohne Bezug auf grünen Tee enthalten.

Für die Visualisierung der Daten werden die Python-Bibliotheken seaborn und Matplotlib verwendet. Darin werden die Daten in Balkendiagrammen visuell auf-

bereitet. Ihr Vorteil besteht darin, dass sich Verhältnisse – und somit auch Gemeinsamkeiten und Unterschiede zwischen den Datensätzen – sehr leicht erkennen lassen.

Zur Verdeutlichung dessen und, um zu zeigen, weshalb die Verwertung der MeSH-Bezeichnungen aus den Metadaten des Korpus in diesem Fall problematisch ist, werden diese den Symptomen der Nature-Untersuchung gegenübergestellt.

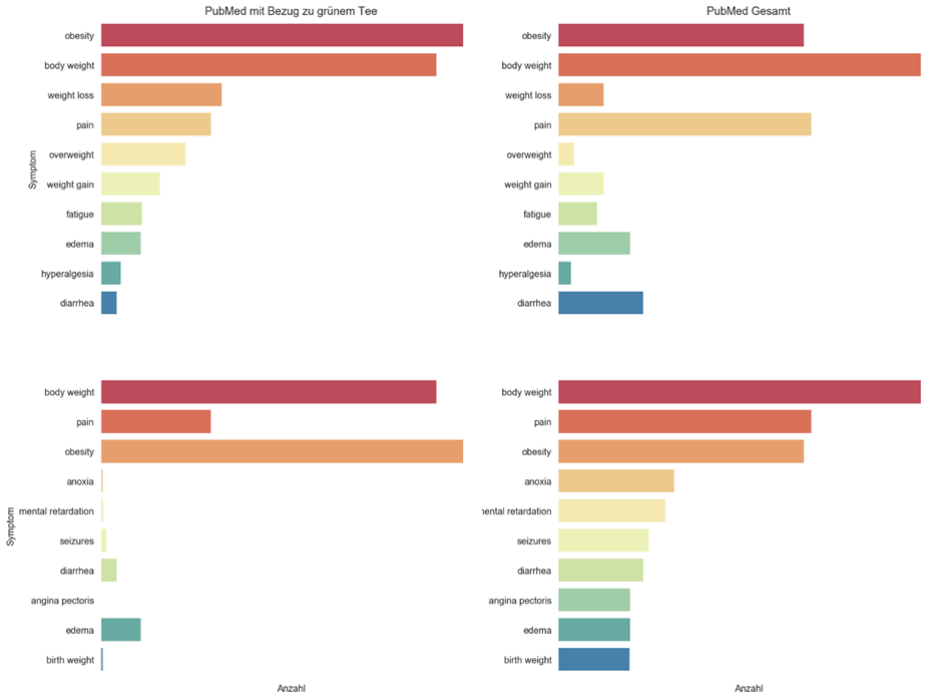
Abbildung 66: Vergleich von MeSH-Bezeichungen mit den Metadaten des Korpus



In Abbildung 66 ist leicht zu erkennen, dass die häufigsten Bezeichnungen aus den Metadaten nur wenig Gegenstücke im PubMed-Datensatz des Nature-Artikels finden. Dafür gibt es zwei Hauptgründe. Zum einen wird im Nature-Artikel eine Unterscheidung zwischen Krankheiten und Symptomen getroffen, die im Metadatensatz so nicht vorliegt. Zum anderen lässt sich aus dem Nature-Datensatz nicht ableiten, welche Ebene der MeSH-Codes die Grundlage der Bezeichnungen ist, während in den Metadaten recht niedrige Ebenen verwendet wurden, sodass spezifischere Bezeichnungen auftreten als im Nature-Datensatz. Somit lassen sich diese beiden Datensätze in dieser Form nicht erkenntnisreich miteinander verbinden.

Besser funktioniert dies mit den Bezeichnungen, welche aus den Abstracts des Korpus ausgezählt wurden.

Abbildung 67: Vergleich extrahierter Symptome aus den Abstracts mit Nature-Daten



In Abbildung 67 ist dargestellt, wie oft bestimmte Symptome relativ in allen PubMed-Daten auftreten. Dabei werden die zehn häufigsten Symptome aus den Abstracts in absteigender Reihenfolge aufgelistet und die Vorkommen aus PubMed insgesamt diesen in gleicher Reihenfolge zugeordnet. Auf der unteren Seite ist diese Zuordnung umgekehrt. Somit richtet sich die Reihenfolge dort nach Vorkommnissen in der gesamten PubMed-Datenbank.

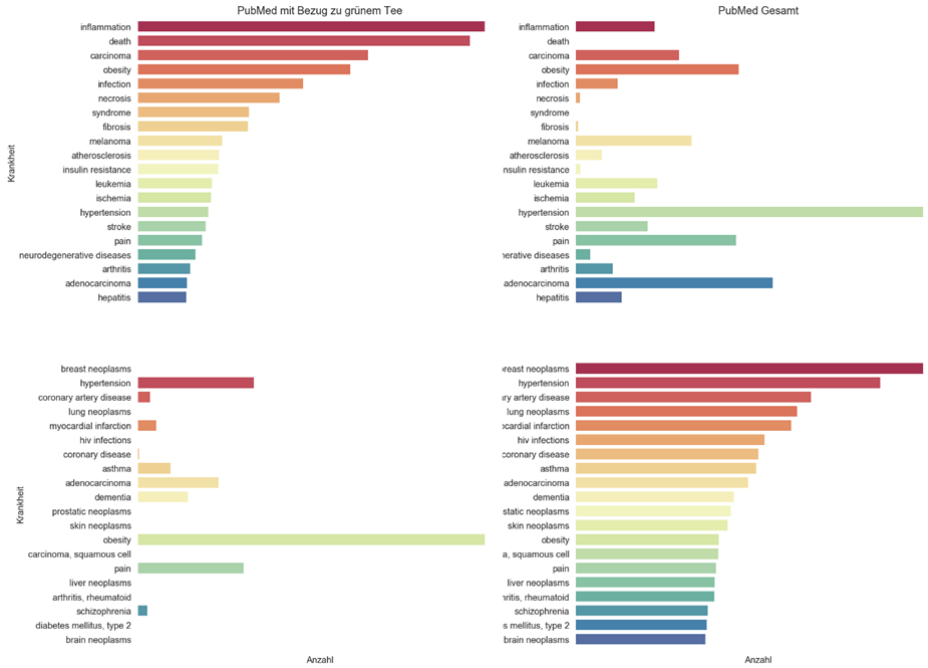
Zunächst lässt sich hier bereits feststellen, dass die Textextraktion der Symptome (im späteren Verlauf auch der Krankheiten) besser mit den Daten des Nature-Artikels verbunden werden kann. Ein möglicher Grund dafür ist, dass grundsätzlich allgemeinere Begriffe in den Abstracts genutzt wurden. Dies scheint nah an dem zu liegen, was die Autoren des Nature-Artikels für die Zusammenfassung der MeSH-Bezeichnungen genutzt haben.

Weiterhin sind einige ähnliche Anteile für verschiedene Symptome zu erkennen. So haben beispielsweise obesity, body weight und pain jeweils hohe Anteile. So-

mit scheinen diese Dinge im Allgemeinen sehr viel in Wissenschaftlichen Arbeiten der Medizin behandelt zu werden. Die Verteilung ist nicht signifikant von der Beschränkung auf das Thema des grünen Tees beeinflusst.

Auf der anderen Seite ist zu sehen, dass es einige Themen gibt, die allgemein sehr viel stärker behandelt werden als im Bezug zu grünem Tee. Dazu zählen beispielsweise anoxia (Sauerstoffmangel), mental retardation (geistige Behinderung) und seizures (Krampfanfälle). Somit ist die Wirkung des grünen Tee-Wirkstoffs in diesen Bereichen entweder nicht hinreichend wissenschaftlich aufgearbeitet oder es wird angenommen, dass keine Behandlung damit möglich ist.

Abbildung 68: Vergleich extrahierter Krankheiten aus den Abstracts mit Nature-Daten



Betrachtet man nun die Krankheiten anstelle von Symptomen, so sind durch eine ähnliche Darstellung ebenfalls einige Auffälligkeiten zu erkennen. Zunächst ist festzuhalten, dass besonders die Bezeichnungen death (Tod) und syndrome (Syndrom) in diesem Kontext kritisch zu betrachten sind. Diese werden im Bezug zu grünem Tee sehr häufig genannt, wohingegen keine Nennungen im Gesamtkontext der PubMed-Datenbank angegeben werden. Dies ist nicht vereinbar.

Naheliegende Begründungen hierfür sind entweder nicht fehlerfreie Textextraktion von Bezeichnungen aus den Abstracts oder eine anders gewählte Abstraktionsebene der MeSH-Bezeichnungen im Datensatz des Nature-Artikels. Ersteres könnte beispielsweise bedeuten, dass das extrahierte Wort death eigentlich für cell death (Zellsterben) steht und dieser Unterschied in der Extraktion verloren gegangen ist. Somit wären die Daten nicht hundertprozentig belastbar. Letzteres könnte bedeuten, dass diese Begriffe zwar korrekt extrahiert worden sind, allerdings nur als Ober- bzw. Unterkategorie der MeSH-Code-Ebene vorkommen, die beim Nature-Artikel verwendet wurde. Auch eine Kombination dieser Gründe ist denkbar.

Ähnliche Gründe könnten erklären, weshalb die im gesamten PubMed sehr häufig untersuchten verschiedenen Neoplasmen in Bezug auf grünen Tee keine Vorkommen erkennen lassen. Dies ist im unteren Teil der Abbildung 68 erkennbar. Weiterhin wird dies dadurch gestützt, dass das gleiche Phänomen auch bei diabetes mellitus, type 2 auftritt, also einer Form der Zuckerkrankheit, obwohl bei dieser bereits Forschung und Behandlung im Zusammenhang mit dem grünen Tee-Extrakt bekannt sind.

Der größte Erkenntnisgehalt lässt sich aus den Teilen generieren, in denen es zwar grafisch deutliche Unterschiede zwischen linker und rechter Darstellung gibt, aber auf jeden Fall auf beiden Seiten Daten vorhanden sind. Damit kann eingeschränkt werden, dass eine (teilweise) Inkompatibilität ein falsches Bild erzeugt. Besonders auffällig sind hierbei inflammation (Entzündung) und necrosis (Absterben von einzelnen oder vielen Zellen). Bei diesen ist eine deutlich stärkere Ausprägung seitens des grünen Tees zu sehen. Ähnliches gilt auch für Fibrose, das krankhafte Wachsen von Bindegewebe, und Insulinresistenz.

Aus diesen Zusammenhängen allein lässt sich – unter anderem bei Berücksichtigung der Datenlage – nicht feststellen, dass der grüne Tee-Extrakt als Heilmittel für diese Krankheiten angesehen werden kann. Fachlich gesehen könnte grüner Tee auch gerade bei diesen Themen besonders stark behandelt worden sein, um vor einer womöglich schädlichen Wirkung von grünem Tee bei diesen Krankheiten zu warnen. Dennoch bietet diese rein datentechnische Untersuchung gute Anhaltspunkte bzw. Indizien, an welchen Stellen es sich lohnt, in der Literatur die Wirkung von grünem Tee genauer zu prüfen, und regt zu ersten Vermutungen in diesem Bereich an.

5.6 Diskussion aller Visualisierungsergebnisse

Bei der MeSH-Code-Exploration im Längsschnitt fällt auf, dass die Nennungen zu grünem Tee seit den 1990er Jahren rasant gestiegen sind. Zwischen den Jahren 1986 und 1989 entwickelte sich eine relativ feste Verteilung, welche sich bis heute gleichmäßig weiterentwickelt. Die Hauptgruppe D *Chemicals and Drugs* wies in den meisten Jahren die größte Anzahl an Nennungen auf und hat im Jahr 2018 im Vergleich zum Vorjahr einen signifikanten Anstieg erfahren. Bei einer vertieften Betrachtung der Hauptgruppe D wurde festgestellt, dass die Nennung der Untergruppe D03 2018 im Verhältnis zu den vorherigen Jahren stark gestiegen ist und alle weiteren Untergruppen prozentual an Anteil verloren haben.

In der Querschnittsbetrachtung zeigt sich ebenfalls die Dominanz der Gruppen *Chemicals and Drugs* (D), *Phenomena and Processes* (G) und *Organisms* (B). Im Vergleich dazu ist erwähnenswert, dass *Diseases* (C) deutlich seltener betrachtet wird.

Auf zweiter MeSH-Code-Ebene stechen folgende Kategorien heraus: Eukaryota (B01), Physiological Phenomena (G07) und Heterocyclic Compounds (D03) sind am stärksten vertreten. Von den Top 10 der MeSH-Codes auf Ebene 2 sind fünf *Chemicals and Drugs* (D), zwei *Phenomena and Processes* (G) und jeweils ein Code *Organisms* (B), *Analytical*, *Diagnostic and Therapeutic Techniques*, *and Equipment* (E) und *Technology*, *Industry*, *and Agriculture* (J) zuzuordnen.

Auf dritter Ebene sind ebenfalls die Hauptgruppen B, G und D in den Top 10 MeSH-Codes vertreten. Wie auf Ebene 2, so ist auch hier die Hauptgruppe D mit den sechs Codes Heterocyclic Compounds, FusedRing (D003.633), Heterocyclic Compounds, 1-Ring (D03.383), Biological Products (D20.215), Pharmacologic Actions (D27.505), Proteins (D12.776), und Specialty Uses of Chemicals (D27.720) am stärksten erwähnt. Darauf folgen zwei Codes der Hauptgruppe G (Diet, Food, and Nutrition (G07.203) und Biochemical Phenomena) und jeweils ein Code zu B und J, Animals (B01.050) und Beverages (J02.200).

Bei der Betrachtung der Hauptgruppe C *Diseases* und der Ermittlung der Top 5 Krankheiten, die laut der Anzahl der MeSH-Codes am stärksten in Relation zu grünem Tee stehen, ergibt sich folgende Rangfolge: Die häufigste Nennung bezieht sich auf Neoplasms, gefolgt von Pathological Conditions, Signs and Symptoms, Digestive System Diseases, Cardiovascular Diseases und Nutritional and Metabolic Diseases. Die stärkste Subgruppe dieser Krankheiten ist Proteins und Enzymes.

Bei der Analyse der MeSH-Codes in der Hauptgruppe A *Anatomy* zeigt sich, dass die dominantesten Kategorien mit den meisten MeSH-Codes Cells, Digestive System, Nervous System, Tissues und Hemic and Immune Systems sind.

Auf der dritten Ebene stechen vor allem die MeSH-Codes Animals (B01.050), Heterocyclic Compounds, Fused-Ring (D03.633), Heterocyclic Compounds, 1-Ring (D03.383) und Proteins (D12.776) am stärksten hervor und zeigen auch die meisten Verbindungen zu anderen MeSH-Codes. Besonders auffällig ist das Dreieck zwischen Animals (B01.050), Heterocyclic Compounds, Fused-Ring (D03.633) und Heterocyclic Compounds, 1-Ring (D03.383).

Betrachtet man die MeSH-Codes der fünften Ebene in Verbindungen mit Krankheiten, also der Hauptgruppe C, zeigt sich, dass ein Großteil der Knoten mit verschiedenen Krebskrankheiten assoziiert sind. Neben den Krebskrankheiten ist auch ein Cluster zu Übergewicht und Fettleibigkeit sichtbar. Auffällig ist weiterhin die starke Verbindung von Vertebrates (B01.050.150.900) beziehungsweise Wirbeltieren zu Carcinoma (C04.557.470.200). Außerdem scheinen Benzopyrans (D03.633.100.150, D03.383.663.283) auch eine bedeutende Rolle zu spielen, da sie als Verbindungspunkt zu verschiedenen Krankheiten stehen.

Wortanalysen zeigen, dass Begriffe wie oxidative beziehungsweise antioxidant, polyphenol, reactive, acid und unter anderem cancer am häufigsten in den Abstracts erwähnt wurden. Dies verweist auf eine lohnende vertiefte Analyse. Die Top 3 Keywords sind demnach polyphenols, oxidative stress und adoptosis. Weiterhin, unabhängig von den MeSH-Kategorien und Keywords, wurden die zehn am häufigsten auftretenden Unigrams ermittelt. Hier wird acid am häufigsten mit grünem Tee assoziiert, gefolgt von polyphenol und antioxidant.

Bei der Differenzierung der Begriffe nach Menschen und Tier, wird deutlich, dass der Begriff oxidative stress fast fünf Mal so oft in Bezug zu Tieren wie in Bezug zu Menschen erwähnt wird. Zusätzlich treten einige Keywords wie Inducible nitric oxid synthase, Eotaxin und Reactive oxygen species in keinerlei Weise in Relation zu Menschen auf, sondern ausschließlich in Bezug zu Tieren. Im Gegensatz dazu sind Begriffe wie inflammation, obesity, antioxidant oder phytochemicals genauso relevant in Bezug zu Menschen.

Die Analyse der meist vorkommenden Noun Phrases auf Basis der Korpus-Metadaten zeigt zunächst nicht überraschend alle zu grünem Tee verwandten Begriffe am dominantesten, wie zum Beispiel: green tea, epigallocatechin gallate egcg, epigallocatechin3gallate egcg und green tea extract. Auch sind weitere Teesorten wie black tea oder oolong tea sichtbar und Begriffe wie oxidative stress

und verschiedene auf Krankheiten bezogene wie prostate cancer, cancer cells, tumor growth oder body weight und insulin resistance.

Insgesamt fallen die „Symptome“ obesity, body weight, weight loss, pain und overweight besonders auf und gehören in allen Betrachtungen zu den am häufigsten genannten Symptomen. Allerdings handelt es sich bei diesen Symptomen ebenfalls um die meistgenannten Bereiche im Nature-Artikel, was darauf schließen lässt, dass diese Themen allgemein sehr häufig in wissenschaftlichen Arbeiten behandelt werden, was wiederum ihre Bedeutung im Kontext Grüntee relativiert.

Bei der Graph-Visualisierung der Krankheiten, lässt sich feststellen, dass inflammation und necrosis im Korpus im Zusammenhang mit grünem Tee oft zusammenhängend betrachtet werden.

Beim Auftreten der „Symptome“ dominieren body weight, weight loss, obesity und weight gain, weshalb sie zu entfernen sind, um eine aussagekräftige Darstellung zu erhalten. Darüber hinaus haben die Symptome pain, fatigue und body weight gemeinsam, dass sie die höchste Anzahl an Verbindungen zu anderen Knoten aufweisen. Die Symptome pain und fatigue haben einen hohen Verbindungscharakter, aber keine starke Verbindung zu bestimmen Symptomen.

Bei der Analyse der Publikationsdaten, der Autorinnen und Autoren sowie des qualitativen Rankings ergibt sich eine differenzierte Verteilung. Während die Anzahl an publizierenden Autoren ein Maximum von über 4.000 erreicht, sind die Extremwerte für die Anzahl an Publikationen ca. 800, für die Anzahl an unterschiedlichen Journals über 400 und für Publikationsländer weniger als 40. Die Verläufe gestalten sich relativ ähnlich. Alle vier Kennzahlen steigen ab 1989 immer steiler an, dabei weist das Diagramm zum Verlauf der Publikationsländer im Verhältnis die größte Schwankung auf. Dies ist auf die niedrige Dimensionierung im Vergleich zu den anderen Kennzahlen zurückzuführen. Allgemein ist festzuhalten, dass in allen Bereichen ein massiver Zuwachs der Publikationen bis 2015 beziehungsweise 2017 festzustellen ist. Das tendenzielle Abfallen aller Kennzahlwerte ab 2018 kann als erstes Indiz für eine einsetzende Sättigung der Themen aus dem Green Tea-Korpus gewertet werden.

Bezüglich der Publikationsquantität der Länder ist eine sehr starke Differenzierung zwischen den Top 5 und allen anderen Ländern festzustellen. Anzunehmen ist, dass sich eine Art Center of Exellence bestehend aus den Publikationsländern USA, England, Niederlande, Deutschland und der Schweiz gebildet hat.

Bei der Auswertung der Publikationspotenz der Autorinnen und Autoren fällt im besonderen Maße der Name Chung S. Yang bzw. C. S. Yang auf. Vergleicht man die Anzahl der Publikationen der Top drei Autoren untereinander, so sticht dieser Autor mit summiert 148 Publikationen im Vergleich zu Y. Hara bzw. Yukihiko Hara mit summiert 122 Publikationen und H. Mukhtar bzw. Hassan Mukhtar mit summiert 114 Publikationen besonders hervor. Allgemein sind Autoren mit asiatischem Namen besonders häufig in der Top 30 Autorenliste vertreten. Dies kann auf die Ursprungsländer des grünen Tees, China bzw. Japan, zurückgeführt werden (vgl. Tanaka, 2012). Betrachtet man die Publikationsverläufe der Top 30 Autorinnen und Autoren, so lässt sich vermuten, dass Publikationen zu diesem Thema von 2002 bis 2018 forciert wurden. Dieses Phänomen kann jedoch auch auf die 2002 in PubMed eingeführte Verwendung von ausgeschriebenen Vornamen anstelle von Kürzeln zurückgeführt werden (vgl. Torvik & Smalheiser, 2009). Teilt man die Analyse der Publikationsverläufe in die Zeiten vor 2003 und ab 2003 auf, so ergeben sich unterschiedliche Verlaufsmuster und Autoren-Rangfolgen für die beiden Intervalle.

In den Verlaufsmustern fällt auf, dass die Extremwerte bei den Top Autoren bis 2003 durchschnittlich etwas tiefer sind als bei denen ab 2003. Damit verbunden sind auch die punktuellen Steigungen: Das Gefälle der Linien des zweiten Intervalls ist steiler als das derjenigen des ersten Intervalls. Dies lässt sich auf die im besonderen Maße gesteigerte Rate an Publikationen ab 2002 zurückführen.

Vergleicht man die Publikationspotenz der Journals, fällt einerseits der Top-Kandidat Journal of Agricultural and Food Chemistry mit 2.665 Publikationen zum zweitplatzierten Journal Food Chemistry mit 1.305 Publikationen auf. Andererseits sind besonders viele Ernährungs- und biochemische Magazine unter den Top 30 vertreten. Dies hängt mit der Zugehörigkeit von grünem Tee zu Ernährungs- beziehungsweise Diätthemen, sowie der Suchparameter für den Korpus zusammen, da auch chemische Abkürzungen aus dem Themenumfeld inkludiert wurden. Die Publikationsverläufe der Top 15 Journals lassen sich in zwei unterschiedliche Gruppen einteilen. Zum einen in die Gruppe der kontinuierlich und über einen längeren Zeitraum publizierenden Journals, zum anderen die Gruppe der kurzfristigeren aber steigernd publizierenden Journals.

Vergleicht man die Verläufe der gesamten Korpus-Publikationen mit den MEDLINE-Zitationen und dem durchschnittlichen H-Index der publizierenden Journals, so stellt man fest, dass die Steigung der Korpus-Publikationen besonders ab 2002 steiler ist als die der MEDLINE-Zitationen. Im Verhältnis betrachtet kon-

kurrieren die Werte ab 2012 wechselnd miteinander. Nimmt man noch den Verlauf des durchschnittlichen H-Indexes aller publizierenden Journals hinzu, so stellt man fest, dass trotz stetig steigender Publikationszahlen, der H-Index nach 2006 steil und stufenweise abfällt. Dies ist, mit Bezug auf den generellen Kennzahlverlauf, als zweites Indiz für eine einsetzende Sättigung der Themen aus dem Green Tea-Korpus zu werten. Dieses Phänomen steht im direkten Zusammenhang zu den zwei Publikationsgruppen der Journals. Die Gruppe der kurzfristig aber steigernd publizierenden Journals hat einen stark negativen Einfluss auf den Verlauf des durchschnittlichen H-Index. Die Begründung dafür liegt in der Berechnungsgrundlage des H-Index. Der plötzliche Anstieg der Anzahl an Publikationen führt in Verbindung mit geringen Zitationen der einzelnen Publikationen zu schlechteren H-Indizes der einzelnen Journals. Zu beachten ist jedoch, dass nicht jedem Journal bzw. jeder Publikation ein H-Index zugeordnet werden konnte. Somit mussten zur Berechnung des durchschnittlichen H-Index alle nicht zugeordneten Publikationen exkludiert werden. Analysiert man das Verhältnis zwischen den Publikationsverläufen der Top 10 Publikationsländer und den durchschnittlichen H-Indizes deren Journals, so lässt sich die bereits angedeutete Sättigung besonders bei den zwei Spitzenreitern (USA und England) beobachten. Platz 3 bis 5 der publikationsstärksten Länder (Niederlande, Deutschland und die Schweiz) weisen einen relativ konstanten bis sogar steigenden durchschnittlichen H-Index auf. Dies lässt auf ein stabiles Forschungs- und Informationsinteresse durch die Bevölkerung dieser Länder schließen.

Setzt man die Ergebnisse der Analysen für die Top 40 publizierten MeSH-Codes sowohl für die nach Publikationen als auch für die nach durchschnittlichem H-Index sortierten Codes miteinander in Beziehung, stellt man fest, dass die Codes der ersten Kategorie in besonderem Maße von US-amerikanischen Publikationen/Journals beeinflusst werden. Zu den Top 10 Codes zählen: Humans; Catechin; Tea, Animals; None (ohne MeSH-Code); Plant Extracts; Antioxidants; Mics; Polyphenols; Flavonoids und Rats. Im Vergleich dazu sind die Top 10 publizierten MeSH-Codes mit dem höchsten durchschnittlichen H-Index: Phenols; Oxidation-Reduction; Flavonids; Mice; Chromatography, High Pressure Liquid; Cells, Cultured; Cell Line; Catechin; Cell Line, Tumor und Signal Transduction. Der markanteste Unterschied zwischen den beiden Gruppen ist, dass die Gruppe der durch den H-Index motivierten Top 10 Codes insgesamt wissenschaftlich signifikantere Benennungen aufweist. Hierdurch wird die wissenschaftliche Gewichtung der einzelnen MeSH-Codes, die am meisten publiziert wurden, erkennbar. Je spezifizierter der MeSH-Code, desto höher ist dessen durchschnittlicher H-Index.

Bezieht man noch die Exkursanalyse zu den Top 40 MeSH-Codes mit mindestens zehn Publikationen und deren Verteilung zu den Top 10 Publikationsländern mit ein, so ergibt sich ein sehr spezialisiertes Trefferbild. Die Top 10 MeSH-Codes sind: Receptor, ErbB-2; Protein Conformation; HIV-1; Nuclear Magnetic Resonance, Biomolecular; Cross-Linking Reagents; Cystein Proteinase Inhibitors; Cyclin-Dependent Kinase Inhibitor p27; Protein Structure, Tertiary; Circular Dichroism und Amyloid. Insgesamt ist eine Verlagerung der Top MeSH-Codes in Bereiche mit besonders medizinischem wissenschaftlich bzw. chemischem Schwerpunkt erkennbar. Hierdurch lassen sich auch länderspezifische Schwerpunkte der Forschung und Publikation identifizieren.

5.6.1 Beeinflussung der Daten und Analysen

Da in PubMed Autorinnen und Autoren mit identischem Namen nicht besonders markiert bzw. unterschieden werden, ist es nicht auszuschließen, dass im Korpus Daten gleichnamiger Autoren undifferenziert vorkommen. Anhand der zu Grunde liegenden Daten ist eine Separierung der Autoren ohne Weiteres nicht möglich. Aus diesem Grund sind die Analysen zu den Autoren möglicherweise weniger aussagekräftig.

Hinzukommt, dass beispielsweise der Autor Chung S. Yang mit kontinuierlichen Publikationen von 2002 bis 2019 sogar als potentester Publizist ermittelt wurde. Jedoch heißt der neunt potenteste Autor C. S. Yang, der kontinuierlich im Zeitintervall von 1992 bis 2001 publizierte. Es ist davon auszugehen, dass es sich hierbei um denselben Autor mit unterschiedlicher Namensschreibweise handelt, da MEDLINE erst ab 2002 die teilweise auch ausgeschriebenen Vornamen der Autoren verwendet. Diese Problematik zu den Autorendaten wurde mehrfach während der Analyse beobachtet und begünstigt eine mögliche Verzerrung der Ergebnisse.

Des Weiteren wurde für die Analysen der Journals nicht konsequent der aktuelle H-Index verwendet. Da der H-Index jedoch unter anderem von der Anzahl der Publikationen abhängig ist und nicht zu allen Journals beziehungsweise Publikationen ein H-Index in den Daten hinterlegt wurde, mussten zur Berechnung des durchschnittlichen H-Index alle Publikationen ohne Angaben exkludiert werden. Aus diesem Grund kann eine eventuelle Kompromittierung der Ergebnisse nicht ausgeschlossen werden.

Ähnlich verhält es sich mit den MeSH-Code-Analysen. Auch die MeSH-Codes sind nicht jeder Publikation im Korpus zugewiesen worden. Durch diesen Sachverhalt kommt es beim Ranking der Top publizierenden Länder im Zusammenspiel mit MeSH-Codes und H-Index zu einer anderen Zusammensetzung als bei den vorherigen Analysen, wodurch auf Platz 10 Indien durch Australien (ursprünglich Platz 16) ersetzt wurde.

Als zusätzliche Problematik kommt die geringe Anzahl an passenden wissenschaftlichen Publikationen zu Publikationsmetadaten sowohl bei PubMed als auch zu den Themengebieten von grünem Tee hinzu. Auf Grund dessen konnten für diese Projektarbeit wissenschaftliche Quellen nur in einem begrenzten Umfang verwendet werden.

5.6.2 Besonderheiten abseits der durchgeführten Analysen

Während der durchgeführten Analysen fiel auf, dass die erste Publikation des Green Tea-Korpus von 1947 in England und die zweite Publikation von 1954 in Russland veröffentlicht wurde. Dieser Sachverhalt ist insofern zu beachten, da aus den Ursprungsländern des grünen Tees, China und Japan, erst 1987 bzw. 1963 die ersten Publikationen in PubMed erfasst wurden (vgl. Tanaka, 2012). Vermutlich wurden bereits früher schon Publikationen zu Themenbereichen des Korpus in anderen nationalen Datenbanken dieser Länder archiviert erfasst, welche PubMed nicht zur Verfügung stehen.

5.6.3 Ausblick

Um bessere Aussagen zur wissenschaftlichen Gewichtung der Autorinnen und Autoren treffen zu können, sollte als nächstes eine Analyse des H-Index der Autoren durchgeführt werden. Des Weiteren wäre eine weiterführende Analyse der MeSH-Codes, auch auf deren hierarchische Struktur sinnvoll. So könnten die MeSH-Codes nach ihren Gruppen abstrahiert und somit zusammengefasst auf deren durchschnittlichen H-Index bewertet werden.

Die MeSH-Codes, besonders im Zusammenhang mit den durchschnittlichen H-Indizes der publizierenden Journals, sind als Themen- und Wissenschaftlichkeitsgröße geeignet. Mit Hilfe dieser Werte können die Publikationen und deren Aussagefähigkeit differenziert werden. Ergänzend zu den durchgeführten Analysen sollten die MeSH-Codes nach ihren Publikationsjahren ausgewertet werden.

Durch diese Ergänzung würde der Faktor Zeit als weitere Kennzahl in die MeSH-Code-Analyse mit einfließen.

6 Fazit

Abschließend ist anzumerken, dass mit den Big-Data-Analyse-Methoden eine MeSH-Code-Exploration durchführbar ist und sich die Quantität der Kategorien sowie eine Entwicklung der relevanten Themen im Zeitverlauf erkennen lässt. Auch lassen sich MeSH-Code Haupt- und Subgruppen im Querschnitt betrachten, ebenso wie das gemeinsame Auftreten von MeSH-Headings. Diese geben zunächst nur einen quantitativen Überblick darüber, in welchen Bereichen sich die Forschung zum Grüntee bewegt. Die inhaltliche Auswertung der Datenvisualisierungen erfordert eingehendere Untersuchungen.

Wortanalysen, N-Grams und Noun Phrases, die mit NLP-Methoden aus den Metadaten ausgelesen und verarbeitet werden, zeigen inhaltliche Schwerpunktsetzungen sowie Konnotationen. Der zusätzlich generierte Datensatz, der rein die mit grünem Tee behandelten Krankheiten und Symptome beinhaltet, ließ weitere spezifische Analysen unabhängig von den MeSH-Kategorien zu. So konnten Krankheiten und Symptome nach Zeiträumen visualisiert, sowie im Rahmen einer Gephi-Visualisierung Verbindungen identifiziert werden.

Die Betrachtung der Autorinnen und Autoren, Publikationen, Journals und Länder anhand der Metainformationen des Korpus ermöglichte die Gewinnung einer Übersicht, in welchen Ländern das Thema grüner Tee eine besondere Bedeutung einnimmt, welche Autoren maßgeblich das Forschungsfeld dominieren und inwieweit das Thema wissenschaftliche Relevanz hat.

Insgesamt betrachtet ist festzuhalten, dass durch die Visualisierung von Publikationsdaten einige Erkenntnisse bestätigt werden, die auch schon auf andere Art und Weise festgestellt wurden. So werden die potenziellen Wirkungen von grünem Tee auf eine Vielzahl von Krankheiten und Symptomen visualisiert, wobei durch die Visualisierungen allein nicht direkt festgestellt werden kann, ob die Wirkungen positiv oder negativ sind.

Dieses zeigt, dass eine Big-Data-Analyse allein ohne das Hinzuziehen von fachlich qualifizierten Personen (z. B. Mediziner und Naturwissenschaftler) nur grobe Hinweise liefert. Dieser Beitrag demonstriert, welche Möglichkeiten der Visualisierung vorhanden sind. Dabei konzentriert sich die Untersuchung auf die häufig auftretenden Wörter oder Codes. Es sei aber auch darauf hingewiesen, dass es mit der Anwendung von Big-Data-Methoden auch möglich ist, Besonderheiten zu entdecken, welche nur sehr selten auftreten.

Literatur

Aitchison, J., & Clarke, S. D. (2004). The Thesaurus: A Historical Viewpoint, with a Look to the Future. *Cataloging & Classification Quarterly*, *37*(3–4), 5–21. https://doi.org/10.1300/J104v37n03_02

BERT GitHub. (2019). *GitHub – Google-Research/bert: TensorFlow Code and Pre-Trained Models for BERT*. https://github.com/google-research/bert

Braun, T., Glänzel, W., & Schubert, A. (2006). A Hirsch-type Index for Journals. *Scientometrics*, *69*(1), 169–173. https://doi.org/10.1007/s11192-006-0147-4

Brückner, M., Westphal, S., Domschke, W., Kucharzik, T., & Lügering, A. (2012). Green Tea Polyphenol Epigallocatechin-3-gallate Shows Therapeutic Antioxidative Effects in a Murine Model of Colitis. *Journal of Crohn's and Colitis*, *6*(2), 226–235. https://doi.org/10.1016/j.crohns.2011.08.012

Buchkremer, R., Demund, A., Ebener, S., Gampfer, F., Jagering, D., Jurgens, A., Klenke, S., Krimpmann, D., Schmank, J., Spiekermann, M., Wahlers, M., & Wiepke, M. (2019). The Application of Artificial Intelligence Technologies as a Substitute for Reading and to Support and Enhance the Authoring of Scientific Review Articles. *IEEE Access*, *7*(c), 65263–65276. https://doi.org/10.1109/ACCESS.2019.2917719

Calgarotto, A. K., Maso, V., Junior, G. C. F., Nowill, A. E., Filho, P. L., Vassallo, J., & Saad, S. T. O. (2018). Antitumor Activities of Quercetin and Green Tea in Xenografts of Human Leukemia HL60 cells. *Scientific Reports*, *8*(1), 3–9. https://doi.org/10.1038/s41598-018-21516-5

Chei, C.-L., Loh, J. K., Soh, A., Yuan, J.-M., & Koh, W.-P. (2018). Coffee, Tea, Caffeine, and Risk of Hypertension: The Singapore Chinese Health Study. *European Journal of Nutrition*, *57*(4), 1333–1342. https://doi.org/10.1007/s00394-017-1412-4

Cichello, S. A., Begg, D. P., Jois, M., & Weisinger, R. S. (2013). Prevention of Diet-Induced Obesity in C57BL/BJ Mice with Addition of 2 % Dietary Green Tea but not With Cocoa or Coffee to a High-Fat Diet. *Mediterranean Journal of Nutrition and Metabolism*, *6*(3), 233–238. https://doi.org/10.1007/s12349-013-0137-z

Dekant, W., Fujii, K., Shibata, E., Morita, O., & Shimotoyodome, A. (2017). Safety Assessment of Green Tea Based Beverages and Dried Green Tea Extracts as Nutritional Supplements. *Toxicology Letters*, *277*(June), 104–108. https://doi.org/10.1016/j.toxlet.2017.06.008

Devlin, J., & Chang, M.-W. (2018). *Google AI Blog: Open Sourcing BERT: State-of-theArt Pre-training for Natural Language Processing*. https://ai.google-blog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Proceedings of NAACL-HLT 2019, 4171–4186. https://www.aclweb.org/anthology/N19-1423.pdf

Ehrnhoefer, D. E., Duennwald, M., Markovic, P., Wacker, J. L., Engemann, S., Roark, M., Legleiter, J., Marsh, J. L., Thompson, L. M., Lindquist, S., Muchowski, P. J., & Wanker, E. E. (2006). Green Tea (−)-Epigallocatechin-gallate Modulates Early Events in Huntingtin Misfolding and Reduces Toxicity in Huntington's Disease Models. *Human Molecular Genetics*, *15*(18), 2743–2751. https://doi.org/10.1093/hmg/ddl210

Google Research. (2019). *Multiclass Text Classification Using BERT and Keras.*

Graham, H. N. (1992). Green Tea Composition, Consumption, and Polyphenol Chemistry. *Preventive Medicine*, *21*(3), 334–350. https://doi.org/10.1016/0091-7435(92)90041-F

Hagiu, A., Attin, T., Schmidlin, P. R., & Ramenzoni, L. L. (2020). Dose-dependent Green Tea Effect on Decrease of Inflammation in Human Oral Gingival Epithelial Keratinocytes: in Vitro Study. *Clinical Oral Investigations*, *24*(7), 2375–2383. https://doi.org/10.1007/s00784-019-03096-4

Hayakawa, S., Oishi, Y., Tanabe, H., Isemura, M., & Suzuki, Y. (2019). Tea, Coffee and Health Benefits. In J.-M. Mérillon, K. G. Ramawat (Hrsg.), *Bioactive Molecules in Food* (S. 991–1047). https://doi.org/10.1007/978-3-319-78030-6_14

Jimeno-Yepes, A., Mork, J. G., Wilkowski, B., Fushman, D. D., & Aronson, A. R. (2012). MEDLINE MeSH indexing: Lessons learned from machine learning and future directions. IHI'12 – Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, 737–741. https://doi.org/10.1145/2110363.2110450

Liu, Z. Q. (2010). Chemical mMthods to Evaluate Antioxidant Ability. *Chemical Reviews*, *110*(10), 5675–5691. https://doi.org/10.1021/cr900302x

Lopes Sakamoto, F., Metzker Pereira Ribeiro, R., Amador Bueno, A., & Oliveira Santos, H. (2019). Psychotropic Effects of L-theanine and its Clinical Properties: From the Management of Anxiety and Stress to a Potential use in Schizophrenia. *Pharmacological Research*, *147*, 104395. https://doi.org/10.1016/j.phrs.2019.104395

Mao, X.-Y., Jin, M.-Z., Chen, J.-F., Zhou, H.-H., & Jin, W.-L. (2018). Live or Let Die: Neuroprotective and Anti-cancer Effects of Nutraceutical Antioxidants. *Pharmacology & Therapeutics*, *183*, 137–151. https://doi.org/10.1016/j.pharmthera.2017.10.012

McEntyre, J., & Lipman, D. (2001). PubMed: Bridging the Information Gap. *CMAJ: Canadian Medical Association Journal = Journal de l'Association Medicale Canadienne*, *164*(9), 1317–1319.

Muthukadan, B. (2019). *Selenium with Python*. https://selenium-python.readthedocs.io/

SCImago. (2019). *Scimago Journal & Country Rank*. Abgerufen am 07.04.2021 von: https://www.scimagojr.com/

Shen, C.-L., Yeh, J. K., Cao, J. J., Tatum, O. L., Dagda, R. Y., & Wang, J.-S. (2010). Synergistic Effects of Green Tea Polyphenols and Alphacalcidol on Chronic Inflammation-induced Bone Loss in Female Rats. *Osteoporosis International*, *21*(11), 1841–1852. https://doi.org/10.1007/s00198-009-1122-8

SpaCy. (2020). *spaCy Industrial-strength Natural Language Processing in Python*. Abgerufen am 07.04.2021 von https://spacy.io/

Stevenson, D. E., & Hurst, R. D. (2007). Polyphenolic Phytochemicals – Just Antioxidants or Much More? *Cellular and Molecular Life Sciences*, *64*(22), 2900–2916. https://doi.org/10.1007/s00018-007-7237-1

Surh, Y. J. (2003). Cancer Chemoprevention with Dietary Phytochemicals. *Nature Reviews Cancer*, *3*(10), 768–780. https://doi.org/10.1038/nrc1189

Tanaka, J. (2012). Japanese Tea Breeding History and the Future Perspective. In L. Chen, Z. Apostolides, Z.-M. Chen (Hrsg.), *Global Tea Breeding* (S. 227–239). Berlin/Heidelberg: Springer. https://doi.org/10.1007/978-3-642-31878-8_6

Textacy. (2019). *textacy: NLP, Before and After spaCy — textacy 0.9.1 Documentation. Abgerufen am 07.04.2021 von https://textacy.readthedocs.io/en/stable/*

Thangapazham, R. L., Passi, N., & Maheshwari, R. K. (2007). Green Tea Polyphenol and Epigallocatechin Gallate Induce Apoptosis and Inhibit Invasion in Human Breast Cancer Cells. *Cancer Biology and Therapy*, *6*(12), 1938–1943. https://doi.org/10.4161/cbt.6.12.4974

Torvik, V. I., & Smalheiser, N. R. (2009). Author Name Disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data*, *3*(3), 1–29. https://doi.org/10.1145/1552303.1552304

US National Library of Medicine. (1996). *Home – PubMed – NCBI*. Abgerufen am 07.04.2021 von https://www.ncbi.nlm.nih.gov/pmc/

Yang, C. S., & Landau, J. M. (2000). Effects of Tea Consumption on Nutrition and Health. *The Journal of Nutrition*, *130*(10), 2409–2412. https://doi.org/10.1093/jn/130.10.2409

Zaveri, N. T. (2006). Green Tea and its Polyphenolic Catechins: Medicinal Uses in Cancer and Noncancer Applications. *Life Sciences*, *78*(18), 2073–2080. https://doi.org/10.1016/j.lfs.2005.12.006

Zhou, X., Menche, J., Barabási, A. L., & Sharma, A. (2014). Human Symptoms-disease Network. *Nature Communications*, *5*(May). https://doi.org/10.1038/ncomms5212

Institut für IT-Management & Digitalisierung
der FOM University of Applied Sciences

FOM Hochschule

FOM – Deutschlands Hochschule für Berufstätige.

Die mit bundesweit über 57.000 Studierenden größte private Hochschule Deutschlands führt seit 1993 Studiengänge für Berufstätige durch, die einen staatlich und international anerkannten Hochschulabschluss (Bachelor/Master) erlangen wollen.

Die FOM ist der anwendungsorientierten Forschung verpflichtet und verfolgt das Ziel, adaptionsfähige Lösungen für betriebliche bzw. wirtschaftsnahe oder gesellschaftliche Problemstellungen zu generieren. Dabei spielt die Verzahnung von Forschung und Lehre eine große Rolle: Kongruent zu den Masterprogrammen sind Institute und KompetenzCentren gegründet worden. Sie geben der Hochschule ein fachliches Profil und eröffnen sowohl Wissenschaftlerinnen und Wissenschaftlern als auch engagierten Studierenden die Gelegenheit, sich aktiv in den Forschungsdiskurs einzubringen.

Weitere Informationen finden Sie unter **fom.de**

ifid

Das ifid Institut für IT-Management & Digitalisierung bündelt Kompetenzen in den Forschungsbereichen Künstliche Intelligenz (KI), Systemwissenschaften, IT-Management und digitale Transformation.

Die Aufgaben des Instituts umfassen Forschung und Entwicklung, Wissenstransfer und Innovationsförderung an der Schnittstelle von Wissenschaft und Praxis. Auch der Transfer von Forschungserkenntnissen in die Lehre spielt eine große Rolle.
Um diese Aufgaben zu erfüllen, setzt die Forschergruppe auf den Einsatz modernster Big Data-Architekturen und KI-Analysesysteme. Es bestehen Kooperationen mit den großen Technologie-Unternehmen und Instituten der Branche.
Die Wissenschaftlerinnen und Wissenschaftler beschäftigen sich insbesondere mit folgenden Feldern:

- Künstliche Intelligenz / Machine Learning / Data Science / Big Data
- Natural Language Processing (NLP)
- Enterprise Architekturen (insbesondere Big Data)
- Einsatz von Blockchain-Technologien
- Digitalisierung von Prozessen
- Integration der Forschung in die Lehre

Weitere Informationen finden Sie unter **fom-ifid.de**