Hochschule Stralsund — Fachbereich Elektrotechnik und Informatik —
Studiengang Medizinisches Informationsmanagement / eHealth — 7. Semester

Bachelor-Thesis

Eine Analyse der Einflussfaktoren der Häufigkeit von Synonymfehlern innerhalb des Record Linkage sowie die Entwicklung eines Modells zur Fehlerminimierung

Name: Aileen Stehn

Matrikelnummer: 19440

Erstgutachter: Prof. Dr. sc. hum. Hans-Heino Ehricke

Externer Zweitgutachter: Dr. rer. med. Martin Bialke

Abgabedatum: 22.02.2024

Danksagung

In Kooperation mit der Unabhängigen Treuhanstelle der Universitätsmedizin Greifswald und unter der fachkundigen Anleitung meines Erstbetreuers Herr Prof. Dr. Ehricke an der Hochschule Stralsund, wurde meine Arbeit verfasst.
Zuallererst möchte ich mich bei Herrn Prof. Dr. Ehricke bedanken, der mich während des gesamten prozess begleitet hat. Herr Ehricke hat mich ermutigt, mein Bestes zu geben und hat stets kritische Einblicke und konstruktives Feedback geliefert.
Ebenso gebührt ein großer Dank der Unabhängigen Treuhandstelle der Universitätsmedizin Greifswald, vertreten durch Herrn Bialke, die mir die Gelegenheit geboten haben, meine Forschung in einem realen Unternehmenskontext durchzuführen. Die Unterstützung, Ressourcen und Einblicke haben meine Arbeit in einem hohen Maß bereichert und mir wertvolle praktische Erfahrungen vermittelt.
Ich möchte an dieser Stelle meine Wertschätzung für Herrn Bialkes fachliche Expertise und sein äußerst engagiertes Mitwirken bei der Entstehung dieser Arbeit ausdrücken. Ein großes Dankeschön geht auch an Herrn Hampf, für seine wertvolle fachliche Unterstützung und seine motivierenden Worte.
Außerdem möchte ich dem Klinischen Krebsregister Mecklenburg-Vorpommern, für die Bereitstellung von Daten und den Austausch über den Goldstandard-Datensatz danken.
Die Zusammenarbeit mit Herrn Bialke und dem gesamten Team der Treuhandstelle war äußerst lehrreich und inspirierend. Ich schätze sehr die Offenheit und die Bereitschaft aller Kolleg:innen, mich bei meiner Forschung zu unterstützen und möchte mich hierfür herzlich bedanken.
Ihre Bereitschaft, wertvolle Zeit zu investieren, ihr konstruktives Feedback und ihre Erkenntnisse haben wesentlich zur Vollendung dieser Arbeit beigetragen.

Abstract

Die Verknüpfung von Datensätzen ermöglicht die Beantwortung komplexer Forschungsfragen, deren Beantwortung anders nicht möglich wäre, jedoch kann die Entstehung von Verknüpfungsfehlern im Rahmen des Verknüpfungsprozess schwerwiegende Konsequenzen für die Forschung haben. Daher ist es notwendig, zu verstehen wodurch die Entstehung von Verknüpfungsfehlern beeinflusst wird, und Strategien zur Fehlerminimierung zu entwickeln.
Diese Arbeit konzentriert sich auf die Entstehung von Verknüpfungsfehlern im Record Linkage prozess sowie die Konsequenzen dieser Fehler und verfolgt das Ziel ein Konzept zur Fehlerminimierung zu entwickeln.

Die Analyse der Einflussfaktoren verdeutlicht, dass die Datenqualität und die technische Umsetzung des Record Linkage Prozess entscheidende Kriterien sind.
Die Darstellung der Konsequenzen von Verknüpfungsfehlern in Bezug auf Selektionsverzerrungen, Fehlklassifizierungen und Informationsverzerrungen unterstreicht die Notwendigkeit der Entwicklung von Strategien zur Fehlerminimierung für Datenquellen mit nicht eindeutigen Identifikatoren, um gravierende Folgen wie eine Unterschätzung eines Krebsrisikos zu verhindern.
Eine Marktanalyse von 11 Record Linkage Lösungen zeigt, dass sich die Vielzahl der Lösungen in ihren Konfigurationsmöglichkeiten ähnelt. Zwei dieser Record Linkage Lösungen, E-PIX und FRIL, werden in dieser Arbeit hinsichtlich ihrer Verknüpfungsqualität unter besonderer Berücksichtigung von Synonymfehlern anhand eines Goldstandard-Datensatzes, evaluiert und auf ihrer Anpassungsfähigkeit hin untersucht, um die Frage zu beantworten, inwieweit bestehende Record Linkage-Lösungen wichtige Verknüpfungskriterien berücksichtigen können.
Beide Lösungen zeigen Stärken bei der Anwendung der Levenshtein-Distanz und weisen in allen Testdurchläufen einen Recall sowie eine Precision im Bereich zwischen 0,99 und 1 auf, zeigen jedoch Schwächen im Umgang mit dynamischen Variablen. Die Integration einer Konfiguration zur Berücksichtigung von Multiple-Value-Feldern erweist sich als äußerst hilfreich bei der Minimierung von Verknüpfungsfehlern. Für beide evaluierten Record Linkage Lösungen kann eine im Rahmen der Testdurchläufe entwickelte Konfiguration als fundierte Basis für Register empfohlen werden.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

BIH	Berlin Institute of Health
DKMS	Deutsche Knochenmarkspenderdatei
DZGs	Deutsche Zentren für Gesundheit
DZHK	Deutsche Zentrum für Herz-Kreislauf-Forschung
EM-Algorithmus	Erwartungsmaximierungs-Algorithmus
E-PIX	Enterprise Identifier Cross-Referencing
EMPI	Enterprise Master Patient
Febrl	Freely Extensible Biomedical Record Linkage
FHIR	Fast Healthcare Interoperability Resources
FRIL	Flexible Record Integration and Linkage
gICS	generic Informed Consent Service
G-Link	Generalized Record Linkage System
gPAS	generic Pseudonym Administration Service
HL7	Health Level 7
IHE	Integrating the Healthcare Enterprise
IMBEI	Institut für Medizinische Biometrie, Epidemiologie und Informatik
IDAT	personenidentifizierende Daten
MII	Medizininformatik-Initiative
MPI	Master Patient Index
NUM	Netzwerk Universitätsmedizin

NYSIIS	New York State Identification and Intelligence System
PID	Personenidentifikator
PPRL	Privacy-Preserving Record Linkage
SAS	Statistical Analysis System
THS	Treuhandstelle
ZIR	Zentrales Identifikationsregister

1. Einleitung

1.1. Motivation und Problemstellung

Das Record Linkage dient der Identifikation von Datensätzen, die zu derselben Identität gehören. Dabei können sich die Datensätze in der selben, oder in verschiedenen Quellen befinden [Dusetzina SB, 2014, S.1].
Record Linkage ist ein wichtiger Prozess in der wissenschaftlichen Forschung, insbesondere im Bereich der Gesundheitsforschung. Die Identifikation und Verknüpfung von Datensätzen aus verschiedenen Quellen ermöglicht es, wertvolle Informationen zu generieren und Forschungsfragen zu beantworten, die mit Daten aus nur einer einzigen Quelle nicht gelöst werden könnten [Intemann et al., 2023, S.1].
Durch die Verknüpfung von Datenquellen entstehen im Rahmen des Gesundheitssektors wertvolle Informationen, die beispielsweise zu einer Verbesserung von Therapieansätzen beitragen, die Identifizierung von Fall- und Kontrollgruppen begünstigen sowie zu einer Verfeinerung von Messwerten beitragen können. Daher kann mit Hilfe des Record Linkage die Versorgungsforschung unterstützt und die Qualität der medizinischen Versorgung verbessert werden [Intemann et al., 2023, Dusetzina SB, 2014, S.1,1].
Das Record Linkage kann darüber hinaus für die Identifizierung von doppelten Einträgen innerhalb eines Datensatzes verwendet werden. In diesem Fall hilft das Record Linkage, die Duplikate zu erkennen, um so sicherzustellen, dass diese die Datenqualität nicht vermindern [Sayers et al., 2015, S.955].
Ein bedeutendes Problem im Record Linkage Prozess ist das Auftreten von Fehlern während des Verknüpfungsprozess, wie dem Homonym- und Synonymfehler.
Der Synonymfehler entsteht, wenn zwei Datensätze zur selben Person zugehörig sind, dies jedoch nicht erkannt wird. Ein Beispiel hierfür wäre eine Veränderung des Nachnamens im Rahmen einer Heirat oder Scheidung, sodass eine Person fälschlicherweise als zwei separate Individuen verwaltet wird.
Bei einem Homonymfehler werden Personen fälschlicherweise als zusammengehörig identifiziert. Ein konkretes Szenario, das solche Fehler verursachen kann, ist die Existenz von Zwillingen, die einen hohen Anteil an denselben personenidentifizierenden Daten (IDAT) aufweisen und daher fälschlicherweise als eine Identität angesehen werden [Hampf et al., 2020, S.2].

Bei einer hohen Anzahl solcher Verknüpfungsfehler können für die Forschung gravierende Folgen wie eine Selektionsverzerrung, Informationsverzerrung, Fehlklassifizierungen oder Messfehler entstehen [Doidge and Harron, 2019, Sariyar et al., 2011, S.2051,684].
Ein konkretes Beispiel für solch gravierenden Auswirkungen ist das DFG-Projekt „Evaluierung eines indirekten Linkage-Ansatzes anhand einer Beispielstudie zum Risiko einer Krebsneuerkrankung und der Krebsmortalität bei Patienten mit Typ-2-Diabetes unter Behandlung mit verschiedenen Antidiabetika“. Hier wurden Daten aus einem Krebsregister verwendet, um Krankenkassendaten der pharmakoepidemiologischen Forschungsbank mit Informationen zum Tumorstadium und zur Todesursache zu ergänzen. Erhöhte Fehlerquoten in diesem Projekt führten dazu, dass das Krebsrisiko unterschätzt wurde [Intemann et al., 2023, S.31,89].
Um von den Vorteilen des Record Linkage zu profitieren und gravierende Folgen wie fehlerhafte Schlussfolgerungen zu vermeiden, ist es daher von besonderer Bedeutung, die Anzahl der Verknüpfungsfehler zu minimieren [Dusetzina SB, 2014, S.1,2].
In einigen Ländern, wie den Skandinavischen Ländern, stehen eindeutige persönliche Identitätsnummern zur Verfügung, die als eindeutiger Identifikator dienen können und daher den Vergleich sowie die Verknüpfung der Daten erleichtern [Harron et al., 2017a, S.5].
In Deutschland sind eindeutige Identifikatoren jedoch aus rechtlichen und ethischen Gründen sowie Datenschutzbedenken noch nicht verfügbar. Daher müssen Quasi-Identifikatoren wie Name, Postleitzahl und Geburtsdatum als Verknüpfungsvariablen verwendet werden, um zwei Datensätze miteinander zu vergleichen. Mit der Verwendung von Quasi-Identifikatoren erhöht sich jedoch das Risiko für die Entstehung von Verknüpfungsfehlern [Weiand, 2022, Intemann et al., 2023, S.4;6,7].
Wird das durch die Verknüpfung entstehende Wissen nicht effektiv genutzt, vernachlässigen Gesundheitsprogramme bedeutende Möglichkeiten, die Daten zur Steigerung der Gesundheit in der Bevölkerung einzusetzen [Dusetzina SB, 2014, S.2].

Die Entwicklung von Strategien zur Fehlerminimierung bei der Verknüpfung von Datenquellen mit nicht eindeutigen Identifikatoren ist von entscheidender Bedeutung, um die Qualität und Aussagekraft der Forschungsergebnisse zu gewährleisten und die, durch die Verknüpfung entstehenden wertvollen Datenressourcen, effektiv zu nutzen.

1.2. Zielsetzung und Arbeitsumfeld

Diese Arbeit konzentriert sich auf die systematische Analyse der Einflussfaktoren von Homonym- und Synonymfehlern im Kontext des Record Linkage und die Entwicklung eines Modells zur Fehlerminimierung.
Die Analyse der Einflussfaktoren erfolgt durch die Betrachtung von Herausforderungen im Record Linkage Prozess in Bezug auf einen Goldstandard-Datensatz und der Datenqualität. Neben der Analyse soll eine Übersicht über verfügbare Record Linkage Lösungen gegeben werden.
Der empirische Teil dieser Arbeit konzentriert sich auf die Record Linkage Lösungen Enterprise Identifier Cross-Referencing (E-PIX) und Flexible Record Integration and Linkage (FRIL), die auf einen Goldstandard-Datensatz angewendet werden. Basierend auf den Erkenntnissen aus der Analyse der Einflussfaktoren werden mit den Record Linkage Lösungen Konfigurationen entwickelt, und analysiert in welchem Rahmen die Record Linkage Lösungen die Einflussfaktoren berücksichtigen können. Die Auswertung der Verknüpfungsergebnisse zielt daher darauf ab, sowohl die Art als auch die Anzahl der Verknüpfungsfehler zu quantifizieren und zu analysieren welche Einflussfaktoren zu den Fehlern führen. Dies ermöglicht Rückschlüsse auf die Effektivität der Record Linkage Lösungen.
Die zentrale Forschungsfrage dieser Arbeit fokussiert sich auf die Frage, inwiefern bestehende Record Linkage Lösungen, die in einer Vielzahl von Projekten eingesetzt werden, die Umsetzung relevanter Kriterien zur Reduzierung von Verknüpfungsfehlern bewältigen können und in welchen Bereichen Optimierungsbedarf besteht.
Ein zusätzliches Forschungsinteresse besteht darin, die Faktoren zu identifizieren, welche die Entstehung von Synonymfehlern im Record Linkage Prozess beeinflussen und deren Konsequenzen aufzuzeigen, wobei ein besonderes Augenmerk auf datenqualitätsbezogenen Aspekten liegt.
Die vorliegende Arbeit entsteht in Kooperation mit der 2014 gegründeten Treuhandstelle (THS). Die THS unterstützt mit ihren fortschrittlichen Software-Lösungen für das sichere und datenschutzkonforme Management von IDAT, Einwilligungen und Pseudonymen, die medizinische und epidemiologische Forschung in Europa [Bialke et al., 2015b, THS, 2023b, S.2].
Ihre zentralen Systemkomponenten, darunter der E-PIX, der generic Informed Consent Service (gICS) und der generic Pseudonym Administration Service (gPAS), werden in nationalen und internationalen Projekten eingesetzt, wobei die Treuhandstelle enge Partnerschaften mit Organisationen wie dem Netzwerk Universitätsmedizin und dem Deutschen Zentrum für Herz-Kreislauf-Forschung pflegt [Bialke et al., 2015b, Bialke et al., 2015a, THS, 2023b, S.3].

1.3. Abgrenzung

Diese Arbeit beabsichtigt die breite Palette an Herausforderungen im Bereich des Record Linkage zu behandeln. Themen wie Datenschutz, ethische Herausforderungen, linguistische Probleme im Kontext internationaler Datensatzverknüpfung und die Skalierbarkeit von Verknüpfungssoftware sind nicht Gegenstand dieser Untersuchung. Vielmehr beschränkt sich die Arbeit auf Verknüpfungsfehler, die während des Record Linkage prozess aufgrund von technischen Herausforderungen und mangelnder Datenqualität entstehen können. Die Arbeit verfolgt nicht das Ziel existierende Record Linkage Lösungen umfassend darzustellen [Kötzschke, 2015]. Stattdessen fokussiert sich die Arbeit darauf, einen Einblick in bestehende Record Linkage Lösungen zu geben.

1.4. Aufbau der Arbeit

Diese Arbeit gliedert sich in insgesamt fünf Kapitel.
Zu Beginn der Arbeit dient das Kapitel „Theoretische Grundlagen“ dazu, einen Rahmen für die nachfolgende Forschung zu schaffen. In diesem Abschnitt werden die Grundlagen des Record Linkage behandelt, einschließlich der Anwendungsgebiete des Linkage-prozess und der Problematik der Verknüpfungsfehler.
Das dritte Kapitel stellt das Record Linkage in der Praxis mit Blick auf spezifische Herausforderungen im Record Linkage Prozess dar. Hierbei werden bekannte Fehlerfälle in Bezug auf die Datenqualität der zu verknüpfenden Daten analysiert sowie die Problematik eines Goldstandard-Datensatzes thematisiert. Außerdem werden die Konsequenzen von Verknüpfungsfehlern dargestellt und ausgewählte Record Linkage Lösungen vorgestellt.
Im vierten Kapitel wird der empirische Teil der Arbeit erläutert. Hierzu gehören die Schritte der Durchführung und die Darstellung der Ergebnisse, bei der die Konfigurationen detailliert gegenübergestellt und ausgewertet werden.
Das folgende Fazit fasst die wichtigsten Erkenntnisse dieser Arbeit zusammen und zieht Schlussfolgerungen aus den Ergebnissen. Außerdem werden die Herausforderungen im Kontext des Record Linkage sowie die erzielten Ergebnisse der vorliegenden Arbeit kritisch hinterfragt. Es werden ebenfalls mögliche Limitierungen der Arbeit und Implikationen für zukünftige Forschung diskutiert.

1.5. Methodisches Vorgehen

Im Zuge einer Literraturrecherche wurde der aktuelle Forschungsstand bezüglich der Herausforderungen beim Record Linkage, den Konsequenzen von Verknüpfungsfehlern und verschiedene Record Linkage Lösungen untersucht.
Die Durchführung dieser Recherche erfolgte mittels verschiedener Datenbanken, darunter Pubmed, CiteSeerX, SpringerLink und IEEEXplore. Zusätzlich zu der Suche in den Datenbanken wurde eine Recherche auf Webseiten der Record Linkage Lösungen durchgeführt, um weitere Informationen zu gewinnen.
Die Suche erfolgte im Zeitraum von Anfang November 2023 bis Mitte November 2023 und beschränkt sich auf Publikationen, die im Zeitraum von 2000-2023 veröffentlicht wurden.

Um spezifische Aspekte des Record Linkage zu beleuchten, wurden die Suchstrings „Record Linkage AND Bias“ sowie „Record Linkage AND Methods“ verwendet. So wurden mit dem Suchstring „Record Linkage AND Bias“ 3.188.520 Treffer in CiteSeerX erlangt, 61.7574 Hits in SpringerLink, 69 Ergebnisse in Pubmed und 7 Hits in IEEEXplore erzielt.
Die Verwendung des Suchstrings „Record Linkage AND Methods“ hat eine gezielte Identifizierung von Arbeiten, die sich auf Methoden zur Durchführung von Record Linkage konzentrieren ermöglicht.
Unter diesem Suchstring wurden 502.932 Hits in CiteSeerX, 176.084 Treffer in SpringerLink, 9.126 Ergebnisse bei Pubmed und 228 Hits in IEEEXplore gefunden.
Um einen genaueren Einblick in das probabilistische Record Linkage zu erhalten, wurde der Suchstring „Fellegi Sunter Model”verwendet. Dieser Suchstring ergab folgende Anzahlen an Treffern: CiteSeerX:2242669, SpringerLink:223, IEEEXplore:469, Pubmed:14.
Die Suche nach Informationen zu verschiedenen Record Linkage Lösungen wurde in den genannten wissenschaftlichen Datenbanken durchgeführt. Hier konnten Veröffentlichungen zu der Mainzelliste, dem LinkKing, Freely Extensible Biomedical Record Linkage (Febrl) und FRIL gefunden werden. Für die weiteren Lösungen wurden Webseiten für die Informationsbeschaffung verwendet. Für das Generalized Record Linkage System (G-Link) und Primat wurden Veröffentlichungen über deren Webseiten gefunden.
Bezüglich der Mainzelliste ergab die Suche mit dem Suchstring „Mainzelliste“ eine geringe Anzahl von nicht mehr als 10 Treffern in den Datenbanken, wobei nur zwei relevante Publikationen ausgewählt wurden.
Die Suche nach der Lösung „LinkKing“ wurde mit dem Suchstring „LinkKing AND Record Linkage“ durchgeführt. Hierbei wurden in Pubmed und SpringerLink mit 20 Treffern die höchste Anzahl an Hits erzielt, während CiteSeerX nur 10 und IEEEXplore nur einen Treffer lieferten.

Die Suche nach Informationen zu „Febrl“ erfolgte durch den Suchstring „Febrl AND Record Linkage“. Die Anzahl der Treffer variiert zwischen 2 und 20, wobei SpringerLink mit 20 die höchste Anzahl und IEEEXplore mit 2 Treffern die niedrigste Anzahl aufwies.
Für „FRIL“ wurden in den Datenbanken 7-20 Hits erzielt.

2. Theoretische Grundlagen

2.1. Anwendungsgebiete und Ziele

Unter dem Begriff „Record Linkage“ versteht man den Prozess der Verknüpfung von Daten, die sich auf dieselbe Identität beziehen, um neue Datenressourcen zu generieren. Record Linkage wird meist unter der Verwendung von IDAT durchgeführt [Weiand, 2022, Harron, 2022, S.4,5;1].
Üblicherweise werden Daten aus zwei oder mehr verschiedener Datenquellen miteinander verknüpft. Des Weiteren kann das Record Linkage dafür verwendet werden, einen Datenbestand auf mögliche Duplikate zu prüfen und sicherzustellen, dass ein Individuum nicht in mehreren separaten Datensätzen vertreten ist [Sariyar et al., 2011, Dusetzina SB, 2014, S.648,1].

Der Einsatz des Record Linkage findet in verschiedenen Sektoren des Gesundheitswesens Anwendung.
Die Verknüpfung von Daten aus verschiedenen Datenquellen und Akteuren erfolgt mit dem Ziel die Versorgungsforschung zu bereichern und wertvolle Einblicke in die Gesundheitszusammenhänge zu gewinnen. Zu den potenziellen Datenquellen und Akteuren gehören Gesundheitsdaten aus Längsschnittkohorten, Biobanken, Pathologiedatenbanken, Gesundheitsdaten aus Arztpraxen und Krankenhäusern, Apothekenaufzeichnungen, Verwaltungsdaten und verschiedene Register, wie beispielsweise Geburts- , Sterbe- , Krebs- und weitere meldepflichtige Krankheitsregister [Dusetzina SB, 2014, Adelaide et al., 2014, Harron et al., 2020, S.2,4,220].
Die Daten können in den Quellen vertikal oder horizonal verteilt sein. Bei vertikaler Datenverteilung enthalten verschiedene Datenquellen unterschiedliche Aspekte derselben Gruppe von Personen. Dies ist der Fall, wenn Patientendaten über verschiedene medizinische Einrichtungen verteilt sind. Diese Art der Verteilung ermöglicht es beispielsweise Therapieoptionen zu erforschen, indem verschiedene Perspektiven kombiniert werden [Intemann et al., 2023, S.12].
Im Gegensatz dazu liegen bei einer horizontalen Datenverteilung verschiedene Personen in unterschiedlichen Datenquellen mit den selben Variablen vor. Dies ist besonders nützlich bei standortübergreifenden medizinischen Forschungsvorhaben, bei denen vergleichbare Daten über verschiedene Gruppen von Individuen hinweg

analysiert werden können [Intemann et al., 2023, S.12].
Record Linkage wird zunehmend in Kohortenstudien zur Bestimmung der Mortalität im Zusammenhang mit Krebskrankheiten eingesetzt. Darüber hinaus ermöglicht die Verknüpfung von Datensätzen die Integration von Daten aus dem Gesundheitssektor mit Informationen aus anderen Sektoren, um detaillierte Untersuchungen spezifischer Gesundheitsprobleme vorzunehmen [Kvalsvig et al., 2019, Moore et al., 2014, S.6,1].

Die Verknüpfung von Daten aus verschiedenen Quellen kann vielfältige Zielsetzungen haben. Die Zusammenführung kann beispielsweise dazu dienen, die Stichprobengröße durch die Anreicherung von weiteren Daten zu erhöhen.
Ein weiterer Grund für den Einsatz des Record Linkage kann darin bestehen, spezifische interessierende Variablen zu erhalten, die erst durch die Verknüpfung entstehen. Auch der Erhalt zusätzlicher interessierender Variablen, die sich aus der Verknüpfung ergeben, ist ein mögliches Ziel [Harron et al., 2020, Intemann et al., 2023, S.220,5].

Die Vorteile die mit der Verknüpfung von medizinischen Datenquellen einhergehen sind vielfältig. So kann beispielsweise der Zeit- und Kostenaufwand erheblich reduziert werden, da bereits vorhandene medizinische Daten verwendet werden können, anstatt notwendige Forschungsdaten neu zu erheben [Harron, 2022, S.1].
Die Stichprobenvergrößerung kann vor allem in der Gesundheitsforschung von großer Bedeutung sein, da eine große Forschungsdaten-Stichprobe zu aussagekräftigeren Ergebnissen führt. Die Repräsentativität der jeweiligen Studie kann daher mit dem Einsatz des Record Linkage gesteigert werden [Moore et al., 2014, S.1].

In der Vergangenheit haben bereits mehrere Verknüpfungen von Datenquellen bewiesen, dass mit Hilfe des Record Linkage relevante Forschungsfragen beantwortet werden können. Ein Beispiel hierfür ist die Verknüpfung des niederländischen Krebsregisters mit der landesweiten niederländischen Pathologiedatenbank, die es ermöglichte, das Risiko und die Prognose von Endometriumkrebes nach einer Behandlung mit Tamoxifen zu untersuchen [Bergman et al., 2000].
Nicht nur die Verknüpfung von Krebsregistern und Pathologiedatenbanken kann dazu beitragen, relevante Forschungsfragen zu beantworten. Auch der Einfluss der Wechselwirkung zwischen Thiazid-Diuretika und genetischen Variationen im Renin-Angiotensin-System auf das Risiko für Diabetes mellitus Typ 2, konnte erforscht werden, indem Apothekenaufzeichnungen mit den Daten aus Biobanken zusammengeführt wurden [Bozkurt et al., 2009].

Diese Beispiele verdeutlichen die Aussage, dass durch den Einsatz von Record Linkage neue, für die medizinische Forschung wertvolle Datenressourcen entstehen können und die Beantwortung von Forschungsfragen ermöglicht wird, die ohne eine Verknüpfung nicht beantwortet werden könnten [Intemann et al., 2023, S.1].

Wenn das Record Linkage als reine Daten-Deduplizierung verwendet wird, ermöglicht dies eine effiziente und sichere Verwaltung von Informationen und trägt zur Genauigkeit und Integrität der Daten bei [Lisbach, 2011, S.10].

2.2. Grundlegender Record Linkage Prozess und Übersicht von Record Linkage Verfahren

2.2.1. Allgemeiner Record Linkage Prozess

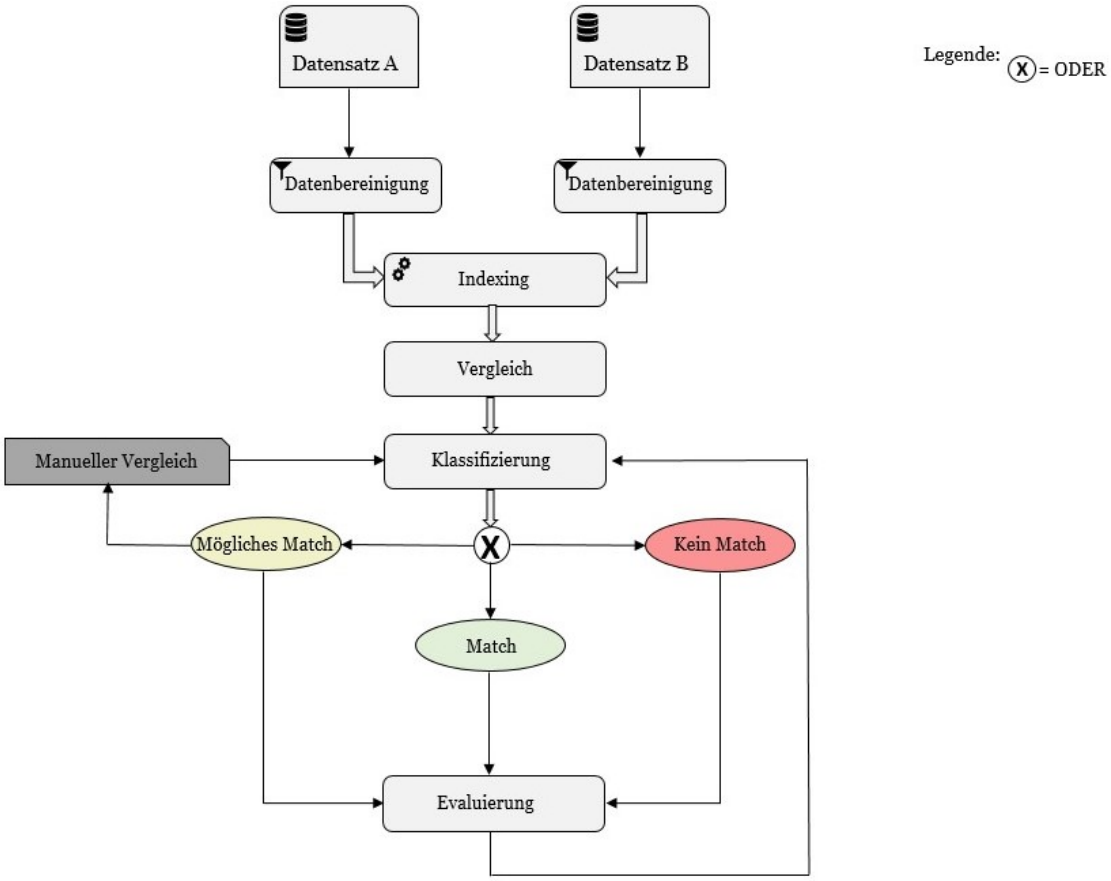

Abbildung 2.1.: Prozess des Record Linkage (adaptiert von [Vatsalan et al., 2017])

Der Prozess des Record Linkage lässt sich in mehrere Schritte unterteilen. Die Auswahl der Identifikatoren (oder Felder/Matching-Variablen) ist der erste Schritt. Hier muss festgelegt werden, welche Quasi-Identifikatoren zur Identifizierung einer Person über mehrere Datenquellen hinweg verwendet werden [Weiand, 2022, S.4,5]. Es gibt verschiedene Kategorien von Matching-Variablen: String Variablen (Name, Adresse), numerische Variablen (Alter) und Kategorische Matching-Variablen (Geschlecht) [Adelaide et al., 2014, S.9].
Die Anzahl der verwendeten Identifikatoren sollte aus Gründen der Datensparsamkeit auf ein Minimum beschränkt werden [DSVGO, 2024, Art.5c Abs.1c].

Bei der Auswahl der Variablen ist zu beachten, dass nicht alle Variablen die gleiche Informationsdichte aufweisen. Einige Identifikatoren, wie etwa die Adresse, besitzen eine geringere Aussagekraft, da die Möglichkeit besteht, dass die Person ihren Wohnort gewechselt hat [Adelaide et al., 2014, Weiand, 2022, S.12,16;3].

Bevor die Daten aus den Datensätzen verglichen werden, sollte als nächster Schritt eine Datenbereinigung erfolgen. Die Daten aus den verschiedenen Quellen werden bei diesem Schritt in die gleiche Struktur und Format überführt, um Fehler innerhalb des Record Linkage prozess aufgrund von ungenauen Daten zu verhindern [Harron et al., 2017a, S.4].

Als Nächstes wird vor allem bei umfangreichen Datensätzen die Technik des sogenannten Indexing angewendet, um die Performance der Verknüpfung zu optimieren. Bei einem direkten Vergleich der Datensätze aus verschiedenen Quellen, würden in der Regel alle Datensätze der einen Quelle mit jedem Datensatz aus der anderen Quelle verglichen werden. Bei umfangreichen Datenquellen wird dies zu einer sehr hohen Anzahl an Vergleichen führen, um das zu verhindern wird häufig das sogenannte Blocking angewendet.
Das Blocking ist ein spezifisches Verfahren des Indexing, hierbei werden die Datensätze in spezifische Blöcke unterteilt und der Abgleich zwischen den Datensätzen erfolgt lediglich innerhalb dieser Blöcke. Ein Beispiel für das Blocking ist die Verwendung der Postleitzahl als Blocking-Variable. Wenn Datensätze in Blöcke basierend auf der Postleitzahl unterteilt werden, werden nur Datensätze mit sehr ähnlichen Postleitzahlen miteinander verglichen, sodass die Anzahl der notwendigen Vergleiche deutlich reduziert wird [Intemann et al., 2023, S.15,16].
Diese Methode des Abgleichs ermöglicht einen effizienteren Verknüpfungsprozess, da nicht alle Datenpaare geprüft werden müssen. Die Auswahl der Blocking-Variablen, nach denen die Datensätze gruppiert werden, ist von entscheidender Bedeutung. Es wird empfohlen, Blocking-Variablen auszuwählen, die eine geringe Fehleranfälligkeit und eine hohe Wertevielfalt aufweisen. Im Idealfall soll das Blocking die Anzahl tatsächlicher Übereinstimmungen erhöhen [Zhu et al., 2009, S.740].

Nachdem mit der Methode des Blockings eine Vorauswahl getroffen wurde, stellt die Verwendung einer geeigneten Verknüpfungsmethode für den Vergleich der Datensätze den nächsten Schritt dar [March et al., 2019, S.646].
Hierbei lassen sich drei Hauptkategorien von Verknüpfungsmethoden unterscheiden: die deterministische Methode, die distanzbasierte Methode und das probabilistische Record Linkage.

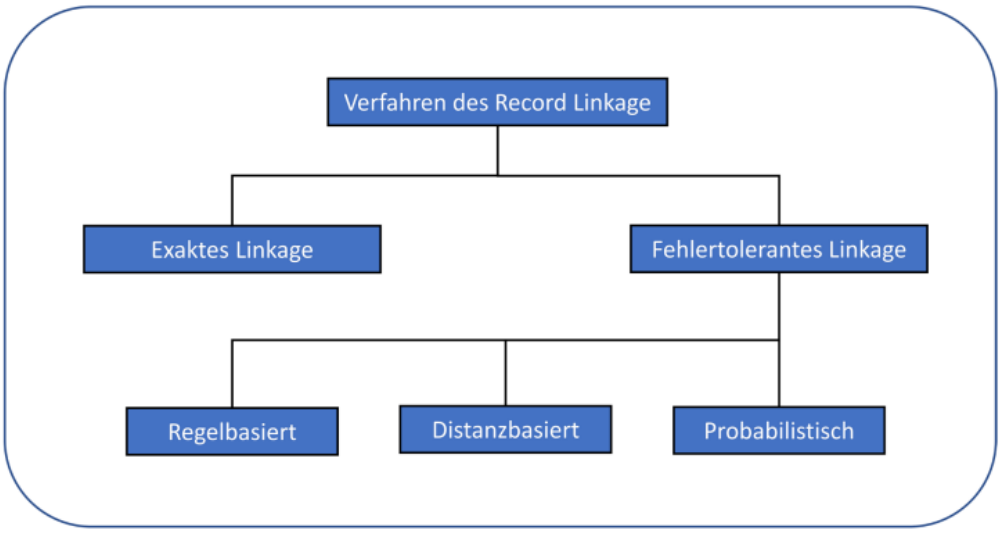

Abbildung 2.2.: Die Verfahren des Record Linkage auf Basis der Publikation [March et al., 2018] (entnommen aus [Intemann et al., 2023, S.16])

Die Methoden weisen Überschneidungen in ihrer Anwendung auf und haben das gemeinsame Ziel, Datenpaare anhand eines Vergleichs gemeinsamer Identifikatoren nach ihrem tatsächlichen Übereinstimmungsstatus zu klassifizieren [Doidge et al., 2020, S.1].
In der Praxis kommen deterministische sowie probabilistische Methoden am häufigsten zum Einsatz [Adelaide et al., 2014, S.8].
Die Auswahl der geeigneten Methode hängt unter anderem von der Qualität der Identifikatoren in den Datenquellen ab. Bei einer hohen Qualität der Identifikatoren kann eine deterministische Datenverknüpfung ausreichend sein, insbesondere wenn ein eindeutiger Identifikator vorhanden ist. Hingegen ist es bei niedriger Datenqualität oft sinnvoller auf eine probabilistische Methode zurückzugreifen, da diese Methode eine gewisse Fehlertoleranz berücksichtigt [Adelaide et al., 2014, S.9].
Es ist üblich Verfahren anzuwenden, die unterschiedliche Ansätze kombinieren [Adelaide et al., 2014, S.8].
Im Rahmen des Abgleichs der Datensätze erfolgt eine systematische Klassifizierung der zu vergleichenden Datenpaare. Wenn zwei Datensätze als zusammengehörig identifiziert werden, erfolgt die Klassifikation als „Match". Falls die Datensätze als nicht zusammengehörig erkannt werden, werden sie als „Kein Match" eingestuft. In Fällen, in denen die Zuordnung nicht eindeutig ist und nicht zweifelsfrei festgestellt werden kann, ob die Datensätze die gleiche Person repräsentieren, werden diese als „Mögliches Match" klassifiziert und unterliegen einer manuellen Prüfung.
Bei einer manuellen Prüfung von potenziellen Übereinstimmungen werden die zu

vergleichenden Datensätze gegenüber gestellt und manuell entschieden, ob eine Übereinstimmung vorliegt [Vatsalan et al., 2017, S.8].

Als letzter Schritt erfolgt die Evaluierung, die dazu dient, die Qualität der Datenverknüpfung zu analysieren. Im Falle identifizierter Schwächen wird eine Anpassung der Konfiguration vorgenommen und es erfolgt eine erneute Klassifizierung der Datenpaare [Vatsalan et al., 2017, S.8].

2.2.2. Deterministisches Record Linkage

Deterministische Verknüpfungsmethoden zeichnen sich insbesondere durch ihre Einfachheit aus, indem sie auf vordefinierten Regelsätzen basieren, die zur Feststellung von Übereinstimmungen oder Nicht-Übereinstimmungen verwendet werden [Doidge et al., 2020, S.1].
Diese Methode beinhaltet einen exakten sowie einen fehlertoleranten Abgleich [Adelaide et al., 2014, S.16].
Bei einem exakten Abgleich werden für jede Matching-Variable k die vorab definierten Regeln so festgelegt, dass sämtliche Werte aller Verknüpfungsvariablen der beiden zu vergleichenden Datensätze übereinstimmen müssen, damit zwei Datensätze als zur gleichen Person zugehörig betrachtet werden [Adelaide et al., 2014, S.8].
In Anbetracht der mangelnden Verfügbarkeit eindeutiger Identifikatoren gestaltet sich ein exakter Abgleich als problematisch, denn bei bei der Verwendung von Quasi-Identifikatoren sollte Raum für potenzielle Fehler gelassen werden [Intemann et al., 2023, S.8].
Das exakte Matching kann verallgemeinert wie folgt ausgedrückt werden [Adelaide et al., 2014, S.11,12]:

$$S = \sum_{i=1}^{k} \text{mit } \boldsymbol{x}_i = \begin{cases} 1, & a_k = b_k \\ 0, & a_k \neq b_k \end{cases}$$

Dabei ist S die Summe der exakt übereinstimmenden Werte aller Matching-Variablen k für ein Vergleichspaar. Anhand dieser Summe wird mit Hilfe eines Schwellenwertes bestimmt, ob die Datensätze zu derselben Person gehören. In der dargestellten Formel stellen a_k sowie b_k die Ausprägungen der jeweilige Variable k dar.
Innerhalb des fehlertoleranten Ansatzes, besteht die Möglichkeit, die vorab definierten Regeln zu lockern, sodass die Übereinstimmungskriterien für einige Variablen weniger streng definiert werden und nicht alle Ausprägungen exakt übereinstimmen müssen, um das Vergleichspaar als Match zu identifizieren [Adelaide et al., 2014,

S.8].
Die deterministische Verknüpfungsmethode weist bei fehleranfälligen Daten eine geringere Verknüpfungsqualität als die probabilistische Verknüpfungsmethode auf. Diese Qualitätsminderung resultiert vor allem aus der signifikanten Anzahl von Nicht-Übereinstimmungen, die bei der Durchführung eines exakten Abgleichs entsteht und zu einer Beeinträchtigung der Identifizierung von echten Übereinstimmungen führen kann [Harron et al., 2017a, S.5,6].

2.2.3. Distanzbasiertes Record Linkage

Während beim deterministischen Record Linkage vorab Regeln definiert werden, in welcher Form die Ausprägungen der Matching-Variablen übereinstimmen müssen, werden bei dem distanzbasierten Record Linkage die Ausprägungen der Matching-Variablen von zwei Datensätzen im Hinblick auf ihre Ähnlichkeit verglichen [Intemann et al., 2023, S.14].
Der Abgleich erfolgt in der Regel mit einem sogenannten Ähnlichkeitsmaß. Die Verwendung eines Ähnlichkeitsmaßes ermöglicht die Identifizierung des Übereinstimmungsgrades eines Datenpaares auf verschiedene Ebenen. Typischerweise gibt es drei verschiedene Ebenen: „identisch“, „nahezu identisch“ und „nicht übereinstimmend“. Der Wert einer Ähnlichkeit liegt bei diesen Techniken zwischen 0 und 1. Bei 1 liegt eine exakte Übereinstimmung der Worte vor [Weiand, 2022, Intemann et al., 2023, S.41,50;15].
Es existieren verschiedene Kategorien von Ähnlichkeitsmaßen. Bei der Klasse der „Edit Distance“ werden Zeichenketten verglichen und gemeinsame Muster identifiziert. Die Ansätze des „Phonetic Encoding“ zielen darauf ab, Wörter in einen Code zu transformieren und ähnlich klingenden Wörtern denselben oder einen ähnlichen Code zuzuordnen. In der dritten Kategorie „Suche mit Varianten“ wie Thesauri, wird zum Abgleich ein Thesaurus verwendet, der verschiedene Schreibmöglichkeiten von bestimmten Wörtern enthält [Lisbach, 2011, S.77,78].

Edit Distance

Als erstes werden spezifische Maße der Kategorie „Edit Distance“ vorgestellt.
Eine bekannte und weit verbreitete Technik ist die *Levenshtein-Distanz*, die auf der Umwandlung einer ersten Zeichenkette in eine zweite Zeichenkette mit einer möglichst minimalen Anzahl von einfügenden, löschenden und ersetzenden Operationen basiert. Die minimale Anzahl an benötigten Operationen bildet die Levenshtein-Distanz [Lisbach, 2011, S.78].

Tier - Tor	**Test - Test**
1. T**i**er = T**o**er	/
2. To**e**r = To**r**	/
Levenshtein-Distanz = 2	Levenshtein-Distanz = 0

Tabelle 2.1.: Beispiele für die Levenshtein-Distanz auf Basis der Publikation [Lisbach, 2011](entnommen aus: eigene Aufnahmen)

Bei der Verwendung der Levenshtein-Distanz muss die maximal zulässige Distanz zwischen zwei Zeichenketten festgelegt werden, bei der von einer Übereinstimmung ausgegangen wird. Diese Wahl beeinflusst die Anzahl der wahren positiven Übereinstimmungen und der falsch Positiven Übereinstimmungen [Lisbach, 2011, S.79].
Ein wichtiger Aspekt ist die Länge der Zeichenkette. Je länger eine Zeichenkette ist, desto größer kann die maximal zulässige Distanz gewählt werden, da bei kürzeren Zeichenketten die Wahrscheinlichkeit geringer ist, dass Unterschiede auf Tippfehler zurückzuführen sind.
Außerdem enthalten zwei längere Zeichenketten einen größeren Anteil an gemeinsamen Informationen als kürzere Namen. Daher werden Abweichungen in längeren Zeichenketten nicht so stark gewichtet, wie in kürzeren Zeichenketten, selbst wenn die Abweichung in beiden Ketten gleich stark ausfällt [Lisbach, 2011, S.79,80].

Die *n-Gramm-Methode* ist ein weiteres Verfahren der Kategorie „Edit Distance“. Bei dieser Technik wird die Zeichenkette in Teilsequenzen zerlegt. Häufig bestehen diese Teilsequenzen aus zwei Zeichen und werden Bigramme genannt.
Wenn zwei Zeichenketten miteinander verglichen werden sollen, erfolgt die Bewertung anhand der Übereinstimmung dieser n-Gramme. Dabei wird analysiert wie viele Bigramme in beiden Zeichenketten vorkommen.
Wenn eine hohe Anzahl an Bigrammen in beiden Zeichenketten übereinstimmen, werden die beiden Zeichenketten als ähnlich betrachtet[Lisbach, 2011, S.81,82].

Bei der *Hamming-Distanz* wird die Anzahl der gleichen Zeichen an denselben Positionen in beiden Zeichenketten gezählt. Die Methode ist sowohl auf gleich lange als auch auf unterschiedlich lange Zeichenketten anwendbar. Wenn zwei unterschiedlich lange Zeichenketten vorliegen, werden die nicht vorhandenen Zeichen in der kürzeren Kette als nicht übereinstimmend gewertet [Weiand, 2022, S.42].
Der Wertebereich der nicht übereinstimmenden Zeichen an bestimmten Positionen reicht von Null bis zur maximalen Anzahl der Zeichen in der längeren Zeichenkette. Um die Anwendung des Ähnlichkeitsmaßes auf unterschiedlich lange Zeichenketten zu ermöglichen, wird in diesem Fall die Anzahl der nicht übereinstimmenden

Zeichen durch die maximale Länge der Zeichenkette geteilt [Weiand, 2022, S.42].

Das *Jaro-Winkler-Ähnlichkeitsmaß* basiert auf der Kombination von verschiedenen Techniken, darunter die Levenshtein-Distanz und die Verwendung von n-Grammen. Es werden mehrere Faktoren berücksichtigt, um die Ähnlichkeit zu ermitteln. So wird die Länge der gemeinsamen Zeichen am Anfang der jeweiligen Zeichenkette mit einbezogen sowie die Anzahl an übereinstimmenden Zeichen in beiden Zeichenketten. Es gibt viele verschiedenen Modifikationen der Distanz [Weiand, 2022, S.43,44].

Phonetische Kodierung

Phonetische Kodierung ist eine Kategorie von Ähnlichkeitsmaßen, bei der die Zeichenketten in phonetische Codes umgewandelt werden und ähnlich klingende Wörter denselben Code erhalten. Ein gängiger Algorithmus für die phonetische Kodierung ist der Soundex-Algorithmus. Der Soundex-Code setzt sich aus dem Anfangsbuchstabe des Wortes sowie drei weiteren Zahlen zusammen. Hierbei bleibt der erste Buchstabe des Wortes unverändert, während die darauf folgenden Buchstaben in Zahlen umgewandelt werden. Jedem Buchstaben wird eine Ziffer nach einer festgelegten Transformationstabelle zugeordnet [Lisbach, 2011, S.83,84].

Buchstabe	**Ziffer**
b f p v	1
c g j k q s x z	2
d t	3
l	4
m n	5
r	6

Tabelle 2.2.: Transformationstabelle von Soundex (reproduziert von [Lisbach, 2011, S.84])

Vokale wie „a“ , „e“ , „i“ , „o“ und „u“ sowie die Buchstaben „w“ , „y“ und „h“ werden ausgelassen. Wenn durch die Transformation zwei gleiche Zahlen nebeneinander stehen, wird der Code angepasst, sodass nur eine der beiden Zahlen in der Zeichenkette verbleibt. Dieser Code wird „Similarity Key“ genannt. Wenn innerhalb des Similarity Keys nach dem Anfangsbuchstabe keine drei weiteren Zahlen vorhanden sind, werden diese Stellen mit Nullen aufgefüllt. Es ist wichtig zu beachten, dass Soundex auf englischen Schreibgewohnheiten basiert [Lisbach, 2011,

S.83,85].
Es gibt jedoch für andere Sprachen angepasste Verfahren, wie zum Beispiel für Deutsch „Kölner Phonetik". Bei diesem Verfahren wurde die Transformation der Buchstaben an die deutsche Sprache angepasst [Boettcher et al., 2014, S.287,289].

Thesauri

Der Einsatz von Thesauri, die verschiedene Schreibvarianten von Worten enthalten, kann ebenfalls als Ähnlichkeitsmaß dienen. Bei dieser Methode wird eine Abfrage an eine Datenbank gesendet, die Einträge von verschiedenen Variationen des gesuchten Namens auswirft.
Es ist jedoch bei diesem Verfahren nahezu unmöglich alle Tippfehlervarianten und unterschiedlichen Schreibweisen in den Thesaurus einzupflegen, sodass der Abgleich in einigen Fällen unsicher sein kann [Lisbach, 2011, S.86,87].

2.2.4. Probabilistisches Record Linkage

Als Nächstes wird die probabilistiche Verknüpfungs-Methode vorgestellt. Fellegi und Sunter entwickelten im Jahre 1969 auf den Vorarbeiten von Newcombe aus dem Jahr 1959 das probabilistische Record Linkage [Intemann et al., 2023, S.15]. Dieses Modell ist von großer Bedeutung in der Datenverknüpfung und basiert auf wahrscheinlichkeits- und statistischen Modellen. Im Vergleich zu deterministischen Verknüpfungsmethoden bieten probabilistische Ansätze oft präzisere und dynamischere Lösungen. Dies ist besonders in Situationen von Vorteil, in denen Quasi-Identifikatoren vorliegen [Zhu et al., 2009, S.739].
Um das Konzept zu verdeutlichen, nehmen wir an, dass zwei Datensätze A und B aus unterschiedlichen Quellen vorliegen, dessen Ausprägungen für eine Matching-Variable als a und b dargestellt werden. In diesem Szenario gehen wir davon aus, dass einige Werte wie a und b in beiden Datensätzen übereinstimmen könnten. Die potenziellen Datenpaare können in zwei Mengen unterteilt werden: in die Menge M (Matched) oder in die Menge U (Unmatched). Die beiden Mengen können wie folgt dargestellt werden [Fellegi and Sunter, 1969, S.3]:

$$M = \{(a, b) | a = b, a \in A, b \in B\}$$
$$U = \{(a, b) | a \neq b, a \in A, b \in B\}$$

Datensätze die tatsächlich zu derselben Person gehören, werden der disjunkten Menge M zugeordnet und echte Nicht-Übereinstimmungen werden der Menge U zugeordnet [Adelaide et al., 2014, S.12].
Ein zentraler Bestandteil des Modells von Fellegi und Sunter sind die feldspezifischen, geschätzten Gewichte für jede Matching-Variable wie beispielsweise Vorname,

Nachname oder Geburtsdatum. Die Idee hinter dieser Gewichtung besteht darin, verschiedene Identifikatoren bzw. Matching-Variablen in den Datensätzen entsprechend ihrer Relevanz zu priorisieren. So erhält beispielsweise das Geburtsdatum in der Regel ein höheres Gewicht als das Geschlecht [Zhu et al., 2009, S.739].
Die Anzahl möglicher Ausprägungen von Matching-Variablen spielt bei der Gewichtung ebenfalls eine Rolle. So bringt beispielsweise die Variable „Name" eine Vielzahl von Ausprägungen mit sich, während das Geschlecht eine begrenztere Variation aufweist. Da die Matching-Variable „Name" mehr Ausprägungen besitzt, ist die Wahrscheinlichkeit, dass zwei verschiedene Identitäten denselben Name teilen, geringer. Daher erhält die Variable „Geschlecht" weniger Gewicht als die Übereinstimmung der Variable „Name" [Intemann et al., 2023, S.15].
Die Berechnung der Gewichte beruht auf der Verwendung von zwei geschätzten bedingten Wahrscheinlichkeiten, die dazu dienen, das Verhältnis zwischen der Wahrscheinlichkeit für eine Übereinstimmung und der Wahrscheinlichkeit für eine Nicht-Übereinstimmung zu bestimmen.
Im Zähler befindet sich die „m-Wahrscheinlichkeit", die die Wahrscheinlichkeit für eine Übereinstimmung in einem spezifischen Feld k unter der Annahme betrachtet, dass das geprüfte Datenpaar tatsächlich eine Übereinstimmung aufweist. Im Nenner des Verhältnisses befindet sich die „u-Wahrscheinlichkeit", die die Wahrscheinlichkeit darstellt, dass das spezifische Feld k trotz der Tatsache, dass das geprüfte Paar keine Übereinstimmung aufweist, dennoch übereinstimmt [Doidge et al., 2020, Adelaide et al., 2014, S.2,12].
Die m-Wahrscheinlichkeit und die u-Wahrscheinlichkeit können wie folgt dargestellt werden [Adelaide et al., 2014, S.12]:

$$m(k) = P([a_k = b_k, a \in A, b \in B] | (a,b) \in M)$$
$$u(k) = P([a_k = b_k, a \in A, b \in B] | (a,b) \in U)$$

Die Berechnung der u-Wahrscheinlichkeit und der m-Wahrscheinlichkeit kann mit verschiedenen Methoden erfolgen. Eine Methode zur Schätzung der m- und u-Wahrscheinlichkeit besteht darin, Informationen über die Wahrscheinlichkeitsverteilung der Variablen und der Wahrscheinlichkeit von Fehlertypen zu nutzen.
Darüber hinaus können etablierte Schätzverfahren wie der Erwartungsmaximierungs-Algorithmus (EM-Algorithmus) und die Maximum-Likelihood-Schätzung genutzt werden [Adelaide et al., 2014, S.17].
Der EM-Algorithmus ist ein iteratives Optimierungsverfahren. Dieser Algorithmus besteht aus zwei Schritten: dem E-Schritt (Erwartungsschritt) und dem M-Schritt (Maximierungsschritt) [Weiand, 2022, S.51].
Im E-Schritt werden Startwerte für die Parameter m und u verwendet. In diesem Schritt erfolgt die Berechnung der erwarteten Werte der Variablen. Diese

erwarteten Werte dienen als Grundlage für die Aktualisierung der Schätzungen in jedem Iterationsschritt. Im anschließenden M-Schritt erfolgt die Aktualisierung der Modellparameter. Basierend auf den erwarteten Werten aus dem E-Schritt werden die Parameter m und u optimiert. Die aktualisierten Parameter werden dann in den nächsten E-Schritt eingesetzt, um erneut die erwarteten Werte der Variablen zu berechnen [Weiand, 2022, S.51].
Dieser Prozess von abwechselnden E- und M-Schritten wird wiederholt, bis eine Konvergenz erreicht ist. Die Konvergenz tritt ein, wenn sich die geschätzten Parameter kaum noch verändern [Weiand, 2022, S.51].
Eine alternative Herangehensweise besteht in der Verwendung des Fuzzy-Algorithmus. Bei diesem Ansatz werden zufällig ausgewählte Datensätze verwendet, wobei jedes Paar verglichen wird und dabei die Anzahl von Übereinstimmungen, Nicht-Übereinstimmungen sowie fehlenden Werten in Bezug auf die Verknüpfungsvariable beobachtet werden. Der Durchschnitt dieser Werte wird berechnet und dient zur Schätzung der Parameter u und m.
Zusätzlich existieren protokollierte Likelihood-Ratios, die verwendet werden können, um das Gewicht der einzelnen Variablen zu bestimmen [Adelaide et al., 2014, S.17,18].
Für jede Variable wird das individuelle Übereinstimmungsgewicht w_{k} berechnet, in das die m-Wahrscheinlichkeit und die u-Wahrscheinlichkeit als grundlegende Faktoren einfließen. Das Gewicht für Variable k kann wie folgt berechnet werden [Adelaide et al., 2014, Weiand, 2022, S.13,51]:

$$\omega_k = \begin{cases} \log_2(\frac{m_k}{u_k}) & \text{, Falls} \quad a_k = b_k \\ \log_2(\frac{1-m_k}{1-u_k}) & \text{, Falls} \quad a_k \neq b_k \end{cases}$$

Das Gewicht eines Datenpaares bildet sich aus der Summe der Gewichte für die vorliegenden Matching-Variablen.
Der Vergleich eines Datenpaares kann bei probabilistischen Verfahren ebenfalls unter der Verwendung von distanzbasierten Methoden erfolgen [Adelaide et al., 2014, S.13].
In diesem Zusammenhang kommt ein Ähnlichkeitsmaß zum Einsatz, welches für jede Matching-Variable analysiert, wie ähnlich sich die Werte eines Datenpaares zueinander verhalten [Adelaide et al., 2014, S.12].
Der Ähnlichkeitswert drückt dabei aus, wie stark die jeweiligen Merkmale der beiden Datensätze übereinstimmen. Aus diesen Werten bildet sich der Vektor y in einem Vergleichsraum (comparison space) Γ. Der Vergleichsraum Γ liegt zwischen

0 und 1. Ein Wert von 1 zeigt an, dass eine exakte Übereinstimmung der Werte vorliegt [Weiand, 2022, Winkler, 2003, S.41,50;1].
Für jede Matching-Variable kann ein Schwellenwert festgelegt werden, der von dem Ähnlichkeitswert der jeweiligen Matching-Variable erreicht oder überschritten werden muss, damit die beiden zu vergleichenden Werte der Matching-Variable als übereinstimmend definiert werden [Sayers et al., 2015, Weiand, 2022, S.960, 52].
Es ergeben sich zwei bedingte Wahrscheinlichkeiten für y [Fellegi and Sunter, 1969]:

$$m(y) = P(y[a,b]|(a,b) \in M)$$
$$u(y) = P(y[a,b]|(a,b) \in U)$$

Im anschließenden Schritt wird das Gesamtgewicht R für das untersuchte Datenpaar berechnet. Dies erfolgt durch die Bestimmung des Verhältnisses der Wahrscheinlichkeit P , dass die Felder a und b aufgrund eines Ähnlichkeitsmaßes zu der Menge der Matches oder der non-Matches gehören. Die Bestimmung des Gesamtgewichts R für ein verglichenes Datenpaar kann wie folgt dargestellt werden [Weiand, 2022, S.50]:

$$R = \frac{P(\mathrm{y} \in \Gamma \mid (\mathrm{a,b}) \in M)}{P(\mathrm{y} \in \Gamma \mid (\mathrm{a,b}) \in U)}$$

Da M und U nicht bekannt sind, muss das Gesamtgewicht auf einem anderen Weg berechnet werden. Eine mögliche Vorgehensweise besteht darin, die m- und die u-Wahrscheinlichkeit jeder Matching-Variable zu verwenden [Weiand, 2022, S.50]. Hierfür kann zunächst das Gewicht, das jeder Variable individuell zugeordnet ist, mit ihrem jeweiligen Ähnlichkeitswert multipliziert werden. Anschließend werden die Produkte derjenigen Matching-Variablen, bei denen der Ähnlichkeitswert den Schwellenwert erreicht oder überschritten hat, addiert.
Parallel dazu erfolgt eine weitere Berechnung, bei der die Produkte derjenigen Variablen addiert werden, bei denen der Schwellenwert für den Ähnlichkeitswert nicht erreicht oder überschritten wurde.
Anschließend wird die Summe der Produkte, die den Schwellenwert erreicht oder überschritten haben, in den Zähler des Verhältnisses eingefügt und die Summe der Produkte, bei denen der Ähnlichkeitswert den Schwellenwert nicht erreicht oder überschritten hat, in den Nenner eingesetzt [Winkler, 2000, S.3].
Ein hohes Gesamtgewicht R weist auf eine höhere Wahrscheinlichkeit für eine Übereinstimmung zwischen zwei Datensätzen hin [Blakely and Salmond, 2002,

S.1248].
Es wird ein oberer Schwellenwert T_μ und ein unterer Schwellenwert T_λ definiert. Wenn das Gesamtgewicht R des Datenpaars den oberen Schwellenwert T_μ erreicht oder überschreitet, wird das Paar als übereinstimmend definiert.
Wenn der Wert den unteren Schwellenwert T_λ erreicht oder unterschreitet, wird das Datenpaar als nicht-übereinstimmend klassifiziert. Befindet sich der R-Wert zwischen dem oberen und unteren Schwellenwert, erfolgt eine manuelle Zuordnung durch einen Menschen, um zu entscheiden, ob die Datensätze zu der selben Person gehören. Es ergeben sich somit folgende Regeln [Schmidtmann et al., 2016, Winkler, 2003, S.2,1]:

Wenn $R \geq T_\mu$	*dann liegt ein Match vor*
Wenn $T_\lambda < R < T_\mu$	*dann liegt ein mögliches Match vor*
Wenn $R \leq T_\lambda$	*dann liegt kein Match vor*

Die Anpassung der Schwellenwerte hat einen Einfluss auf die Anzahl der Datenpaare die als übereinstimmend und nicht-übereinstimmend erkannt werden. Bei einer Erhöhung der Schwellenwerte erhöht sich die Anzahl an erkannten Nicht-Übereinstimmungen, da der obere Schwellenwert seltener erreicht wird und sich somit die Wahrscheinlichkeit einer Nicht-Übereinstimmung erhöht.
Werden die Schwellenwerte gesenkt, wird der obere Schwellenwert schneller erreicht, was zu einer Zunahme der als übereinstimmend identifizierten Datenpaare und einer Verringerung der erkannten Nicht-Übereinstimmungen führt [Weiand, 2022, Adelaide et al., 2014, S.37,27].
Die Auswahl und Festlegung der Schwellenwerte sollte so erfolgen, dass ein ausgewogenes Verhältnis zwischen falsch positiven und falsch negativen Ergebnissen gewährleistet wird. Gleichzeitig sollte angestrebt werden, die Anzahl der Ergebnisse, die zwischen den Schwellenwerten liegen und daher einer manuellen Prüfung unterliegen, zu minimieren. Die Festlegung dieser Schwellenwerte orientiert sich in der Praxis oft an Erfahrungen und bewährten Verfahren [Adelaide et al., 2014, S.18].

2.3. Verknüpfungsfehler

Eine Herausforderung beim Record Linkage besteht darin festzustellen, ob zwei Datensätze tatsächlich zu derselben Identität gehören oder nicht [Adelaide et al., 2014, S.8].
Während des Record Linkage prozess ergeben sich, im Rahmen der Zusammenführung von Datenpaaren, vier unterschiedliche Fälle.

Die Datenpaare, die der Matching-Algorithmus als zusammengehörig identifiziert, werden auch als Treffer, Hits, Matches oder Positives bezeichnet. Datenpaare bei denen davon ausgegangen wird, dass diese nicht zusammengehören, werden als Non-Matches oder Negatives bezeichnet. Es können die folgenden vier Fälle unterschieden werden [Lisbach, 2011, S.12]:

- True Positive = Datenpaare die als Match klassifiziert worden sind und zur selben Identität gehören
- False Positive = Datenpaare die als Match klassifiziert worden sind und zu verschiedenen Identität gehören
- True Negative = Datenpaare die als Non-Match klassifiziert worden sind und zu verschiedenen Identität gehören
- False Negatives = Datenpaare die als Non-Match klassifiziert worden sind und zur selben Identität gehören

Es können sich daher falsche und fehlende Übereinstimmungen beim Record Linkage ergeben [Harron et al., 2017b, S.1700].

		True match status	
		Match (pair from same individual)	**Non-match (pair from different individuals)**
Assigned (predicted) link status	**Link**	True link (true positive) *a*	False link (false positive) *b*
	Non-link	Missed link (false negative) *c*	True non-link (true negative) *d*

Abbildung 2.3.: Klassifikation (entnommen aus: [Doidge et al., 2020])

Das Ziel des Record Linkage besteht darin True Positives und True Negatives zu finden, da diese Fälle korrekte Entscheidungen widerspiegeln. Die anderen beiden Fälle, False Negative und False Positive, stellen wie in Kapitel 1.1 erläutert den Homonym- und Synonymfehler dar [Lisbach, 2011, S.12].
Das Auftreten dieser Verknüpfungsfehler kann in mathematischen Ausdrücken dargestellt werden. Angenommen A1 stellt die Entscheidung „matched“ (zusammengehörig) dar, A2 steht für „possible matched“ (möglicherweise zusammengehörig) und A3 stellt die Entscheidung „unmatchted“ (nicht zusammengehörig) dar, dann lässt sich der Homonym- und Synonymfehler wie folgt darstellen [Fellegi and Sunter, 1969, S.5]:

$$P(A1|U)=\sum_{y} u(y)\,P(A1|y)$$
$$P(A3|M)=\sum_{y} m(y)P(A3|y)$$

Die beiden mathematischen Ausdrücke stellen die bedingten Wahrscheinlichkeiten für die Entscheidungen (A1, A3) unter verschiedenen Bedingungen (U, M) dar.
P(A1|U) präsentiert die Wahrscheinlichkeit, dass die Entscheidung „matched“ (A1) ist, gegeben dass U (Unmatched) eingetreten ist. Das heißt es wird die Wahrscheinlichkeit gemessen, dass ein Datenpaar als zusammengehörig identifiziert wird, obwohl es tatsächlich zu der disjunkten Menge U gehört, die Datenpaare enthält, die nicht zusammengehören.
P(A3|M) beschreibt die Wahrscheinlichkeit, dass die Entscheidung „unmatched“ (A3) ist, gegeben dass M (Match) eingetreten ist [Fellegi and Sunter, 1969, S.5].
Trotz ständiger Verbesserung der Verknüpfungsmethoden bleibt die grundlegende Herausforderung bestehen, ein ausgewogenes Verhältnis zwischen der Anzahl falsch positiver und falsch negativer Matches zu finden. Diese Herausforderung erfordert die sorgfältige Abwägung verschiedener Faktoren, insbesondere bei der Festlegung von Schwellenwerten.
Das zuvor erläuterte DFG-Projekt in Kapitel 1.1 verdeutlicht die Relevanz dieser Herausforderung durch die Unterschätzung des Krebsrisikos infolge von einer Vielzahl an Verknüpfungsfehlern [Adelaide et al., 2014, Intemann et al., 2023, S.8;31,89].
Die Verknüpfungsfehler lassen sich teilweise auf technische Ursachen sowie auf Probleme in Bezug auf die Datenqualität zurückführen. Vor allem bei größeren Datensätzen entstehen häufig falsche Übereinstimmungen, mit denen eine Selektions- oder Informationsverzerrung einhergehen kann [Kvalsvig et al., 2019, Doidge and Harron, 2019, S.9;2051,2053].
Die Untersuchung und Identifikation dieser Fehlerquellen wird in den nachfolgenden Kapiteln detailliert thematisiert.

Um das Ausmaß der Verknüpfungsfehler in Bezug auf die Verknüpfungsqualität eines Record Linkage Prozess zu bewerten, eignen sich Metriken wie Recall, Precision und das F-Maß.
Die Precision gibt den Anteil der als zusammengehörig identifizierten Datensätze, die tatsächlich zusammengehören wieder. Falsch positive Ergebnisse können die Precision daher negativ beeinflussen. Die Precision wird durch folgende Formel ermittelt[Derczynski, 2016, S.262]:

$$P = \frac{|true\ positives|}{|true\ positives| + |false\ positives|}$$

Der Recall gibt Aufschluss darüber, wie viele der tatsächlich zusammengehörigen Datensätze korrekt identifiziert wurden. Folglich können falsch negative Ergebnisse den Recall verringern. Der Recall wird durch folgende Formel berechnet[Derczynski, 2016, S.262]:

$$R = \frac{|true\ positives|}{|true\ positives| + |false\ negatives|}$$

Das F-Maß kombiniert Recall und Precisison und stellt somit ein Gleichgewicht zwischen beiden Metriken dar. Diese Metrik wird als harmonisches Mittel beschrieben und lässt sich durch folgende Formel berechnen[Derczynski, 2016, S.262]:

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Ein F1-Score von 1 signalisiert ein perfektes Gleichgewicht zwischen Precision und Recall, indem keine falsch positiven oder falsch negativen Ergebnisse vorliegen[Derczynski, 2016, S.262].

3. Record Linkage in der Praxis

3.1. Herausforderungen

3.1.1. Fehlerfälle Datenqualität

Im Prozess des Record Linkage können Verknüpfungsfehler auftreten, die durch verschiedene Faktoren verursacht werden. Eine der wesentlichen Ursachen für die Entstehung von Verknüpfungsfehlern ist die Qualität der zu verknüpfenden Daten. Fehlerhafte Werte und fehlende Werte der Matching-Variablen können zu einer erhöhten Anzahl von Verknüpfungsfehlern beitragen, da diese die Identifikation von übereinstimmenden Datensätzen erschweren. Daher ist die Verknüpfungsqualität unter anderem von der Ausgangsqualität der Matching-Variablen abhängig [March et al., 2019, S.647].
Untersuchungen haben gezeigt, dass die Fehlerrate in Krankheitsregistern zwischen 4% und 15% für Variablen wie Nachname, Vorname, Postleitzahl und Geburtsdatum liegen kann. Die Rate fehlender Werte kann zwischen 0% und 9% variieren [Zhu et al., 2015, S.80].
Der Einfluss auf die Entstehung von Fehlern variiert je nach Matching-Variable. Hierbei ist die Anzahl und Verteilung der möglichen Ausprägungen entscheidend [Weiand, 2022, S.11].

Eindeutige Identifikatoren sind in Deutschland lediglich in Studien oder Registern zu finden, diese sind jedoch nicht über verschiedene Einrichtungen hinweg eindeutig, was ihre Nutzung für die Verknüpfung mit andere Datenquellen verhindert. Potenzielle eindeutige Identifikatoren wie die Sozialversichertennummer oder die Steueridentifikationsnummer könnten in Deutschland zwar als eindeutige Identifikatoren dienen, dürfen jedoch nur für spezifische, eng begrenzte Zwecke verwendet werden und sind nicht in allen Datenquellen verfügbar, was ihre Anwendbarkeit einschränkt. Daher werden in Deutschland die Quasi-Identifikatoren für das Record Linkage verwendet[Intemann et al., 2023, S.7,13].
Hinsichtlich der verwendeten Quasi-Identifikatoren ist zu beachten, dass sie eine Vielzahl von Fehlern aufweisen können, die über das Maß eines einzelnen eindeutigen Identifikators hinausgehen. Folglich beeinträchtigen Quasi-Identifikatoren die Qualität der Verknüpfung von Daten.

Ein bedeutendes Problem bei der Verwendung mehrerer Datenquellen und -systemen liegt in der unterschiedlichen Erfassung, Speicherung, und Aktualisierung der Daten. Dies führt dazu, dass die Daten teilweise in unterschiedlichen Formaten

und Kodierungen gespeichert werden [Adelaide et al., 2014, S.10].
Zum Beispiel kann die Kodierung der kategorialen Variable „Geschlecht“ in Datenquellen unterschiedlich ausgeführt werden, etwa „w/m“ in einer Quelle und „f/m“ in einer anderen [Harron et al., 2017a, S.4].
Ein weiteres Beispiel für diese Problematik ist die Erfassung von Datumsangaben. Ein Datum kann in verschiedenen Formaten dokumentiert werden, sowohl in englischer als auch in deutscher Variante, sodass sich der Tag und der Monat nicht mehr an der gleichen Position im Geburtsdatum befindet. Darüber hinaus können Tag, Monat und Jahr durch unterschiedliche Trennzeichen wie Kommata, Leerzeichen oder Bindestriche voneinander getrennt sein [Tromp et al., 2006, Sayers et al., 2015, S.1,958].
Verschiedene Dokumentationsformate sind auch für geografische Standorte, wie beispielsweise Adressen, relevant. Diese können ebenfalls in unterschiedlichen Formen erfasst werden, sei es in ausgeschriebener Form oder in Form von Abkürzen wie ZIP-Codes [Dusetzina SB, 2014, S.30].
Die Verwendung von Präfixen wie „Frau/Dr./Prof.“ und Suffixen wie „Senior“ kann ebenfalls variieren und zu weiteren Inkonsistenzen in den Daten führen [Sayers et al., 2015, S.958].

Auch die Dateneingabe in die Datenquellen A und B erfolgt oft auf unterschiedlicher Art und Weise. So können verschiedene Variationen der einzutragenden Wörter entstehen. Das Auftreten verschiedener Variationen ist abhängig davon, ob die Wörter abgeschrieben, telefonisch übermittelt, oder in ein Diktiergerät gesprochen werden. Dies führt zu Worten mit unterschiedlichen Schreibweisen und kann bei einem Vergleich zweier Werte dazu führen, dass diese als nicht zusammengehörig angesehen werden, obwohl sie tatsächlich zusammengehören [Lisbach, 2011, S.21].
Solche Fehler in der Dateneingabe können ebenfalls auftreten, wenn beispielsweise ein französischer Name mit Diakritika auf einer deutschen Tastatur eingegeben wird. Dabei kann versehentlich die Ausrichtung der Diakritika verwechselt, ähnliche Buchstaben benutzt, oder Buchstaben ganz ausgelassen werden [Lisbach, 2011, S.72,73].
Ein weiteres Problem in Bezug auf die Dateneingabe entsteht, wenn Namen von Personen aus anderen Kulturkreisen diktiert werden. Häufig werden diese Namen falsch ausgesprochen, weil die Menschen mit der Kultur nicht vertraut sind [Weiand, 2022, S.20].
Es ist wichtig, zwischen Variationen und Schreibfehlern zu unterscheiden. Die verschiedenen Variationen sind keine Schreibfehler, da sie auf die Art der Datenübermittlung zurückzuführen sind[Lisbach, 2011, S.71].
Es lässt sich daher sagen, dass die Fehler in den zu vergleichenden Datensätzen

unter anderem von der Datenerfassung und der Dateneingabe abhängig sind [Tromp et al., 2006, S.1].

Fehler in den Datensätzen variieren nach Variablentypen, wobei Schreibfehler insbesondere bei String-Variablen häufig auftreten. Schreibfehler können auf verschiedene Arten entstehen, darunter Einfügungen, bei denen zusätzliche Buchstaben in ein Wort eingesetzt werden, Transpositionen, bei denen Buchstaben an falsche Stellen verschoben werden und Löschungen, bei denen versehentlich Zeichen entfernt oder ersetzt werden [Adelaide et al., 2014, S.10].
Diese Tippfehler können auf motorischen Fehlern, Problemen mit der optischen Zeichenerkennung, Orthographie-Fehlern oder auf Tastaturproblemen, wie beispielsweise einer nicht ergonomischen „Typing Distance", beruhen [Lisbach, 2011, Schuster, 2002, S.71-73,5].
Bei String-Variablen wie dem Vorname können sogenannte „Multiple-Value-Felder" vorliegen. Bei diesen Feldern befinden sich mehrere Zeichenketten in einem Feld, die in ihrer Reihenfolge vertauscht vorliegen können, wie im Beispiel „Anna Lena" / „Lena Anna" [Sayers et al., 2015, 958].
Es ist bekannt, dass vor allem bei demografischen Informationen wie der Adresse häufig Tippfehler auftreten und dass Fehler bei der Dateneingabe häufig im späteren Verlauf des Tages entstehen [Tromp et al., 2006, Dusetzina SB, 2014, S.1,29].

Bei den numerischen Variablen können zusätzlich zu den Tippfehlern Zahlendreher zu einem Fehler führen [Hampf et al., 2020, S.2].
Handelt es sich um Messwerte, können auch Rundungsfehler auftreten, wenn keine präzisen Angaben vorliegen [Adelaide et al., 2014, S.10].
Kategoriale Variablen hingegen weisen in der Regel weniger Eingabefehler auf, da diese oft durch kurze Codes, wie beispielsweise „w" und „m" für das Geschlecht, repräsentiert werden. Dennoch können Unstimmigkeiten in der Zuordnung solcher kategorialen Variablen auftreten [Adelaide et al., 2014, S.10].

Dynamische Matching-Variablen, die sich im Laufe der Zeit verändern können, stellen eine weitere Herausforderung dar [Harron et al., 2017b, S.1700].
Der Vorname ändert sich in der Regel selten, während der Nachname aufgrund von Heirat oder Scheidung wechseln kann [Kvalsvig et al., 2019, S.9].
Auch die Adresse von Personen kann sich aufgrund eines Wechsels des Wohnortes häufiger ändern. Solche Änderungen können dazu führen, dass in Datensätzen unterschiedliche Werte in den Matching-Variablen auftreten, was wiederum zu falschen Zuordnungen führen kann [Harron et al., 2017b, S.1700].

Neben fehlerhaften Werten können auch fehlende Werte in den Matching-Variablen die Datenqualität beeinträchtigen [Harron et al., 2017b, S.1700].
Dies kann auf Unvollständigkeit der Daten zurückzuführen sein oder darauf, dass eine Person keine Ausprägung einer bestimmten Variable zugeordnet werden kann, etwa wenn jemand angibt, kein Geschlecht zu haben und daher keine Wertzuweisung in dieser Variable erfolgt [Harron et al., 2017a, S.2].

Die genannten Fehler, die die Datenqualität beeinträchtigen, lassen sich in zwei Hauptkategorien einteilen: zufällige Fehler und systematische Fehler. Wenn aus zufälligen Fehlern Verknüpfungsfehler entstehen, dann können diese leichter statistisch angepasst werden, als Daten die nicht auf Zufälligkeit beruhen [Doidge and Harron, 2019, S.2053].
Zu den systematischen Fehlern gehören beispielsweise die falsche Schreibweise von Nachnamen von Personen aus anderen Kulturkreisen, unterschiedliche Kodierungen, Namensänderung durch Heirat oder Scheidung und Adressänderungen aufgrund eines Wohnortwechsels [Adelaide et al., 2014, Kvalsvig et al., 2019, S.10,9].
Zufällige Fehler hingegen beinhalten Tippfehler wie Einfügungen, Ersetzungen und Zahlendreher bei Namen, Geburtsdaten oder Postleitzahlen. Zufällige Fehler, die nicht vom Identifikationswert ausgehen und somit in jedem anderen Datensatz auch enthalten sein können, weisen trotz ihrer zufälligen Natur häufig gewisse Regelmäßigkeiten in den Fehlern auf [Adelaide et al., 2014, S.10].
Diese Regelmäßigkeiten können bei der Entwicklung von Matching-Verfahren berücksichtigt werden.

3.1.2. Goldstandard-Datensätze

Im Rahmen des Record Linkage wird unter einem Goldstandard ein Datensatz verstanden, bei dem sämtliche Duplikate bekannt sind und dieser daher als Referenz zur Evaluierung der Leistung von Record Linkage Algorithmen dienen kann. Die Feinjustierung von Matching-Konfigurationen, mit dem Ziel eine verbesserte Qualität der Verknüpfungsergebnisse zu erzielen, kann ebenfalls durch die Nutzung eines Goldstandard-Datensatzes erfolgen.
Da zum Zeitpunkt der Erstellung dieser Arbeit und nach intensiver Rechere in den Literaturdatenbanken Pubmed, CiteSeerX, SpringerLink und IEEEXplore derzeit ein einheitlicher Testdatensatz für Record Linkage Lösungen nicht abschließend definiert ist, ermöglicht die Erstellung eines Goldstandard-Datensatzes eine Beurteilung, inwieweit der Algorithmus Duplikate erkennt.
Ein Goldstandard-Datensatz zeichnet sich durch Repräsentativität aus, insbesondere in Bezug auf die Verteilung der Qualität von Variablen. Die Evaluierung eines Algorithmus anhand dieses Referenzdatensatzes ermöglicht die Bestimmung von Sensitivität, Spezifität und Fehlercharakteristiken des Algorithmus

[Harron et al., 2017b, Doidge et al., 2020, Sayers et al., 2015, S.1702,8,963].
Allerdings stellt die Beschaffung reiner Goldstandard-Datensätze in der Praxis ein bedeutendes Problem dar, da tatsächliche Übereinstimmungen in realen Datensätzen nie absolut bekannt sind. Daher werden in der Praxis verschiedene Methoden angewandt, um möglichst präzise Goldstandard-Datensätze zu erstellen[Weiand, 2022, S.6].

Eine Möglichkeit besteht darin, die Verknüpfung von Datenquellen mit eindeutigen Kennungen als Goldstandard zu verwenden [Harron et al., 2017b, S.1702].
Jedoch können selbst in Datensätzen mit eindeutigen Identifikatoren Fehler auftreten, die zu Verknüpfungsfehlern führen können [Harron, 2022, S.1].
Ein alternativer Ansatz ist die Nutzung synthetisch generierter Daten. Bei künstlich hergestellten Datensätze ist die Validität der Klassifizierung gegeben, jedoch muss die Charakteristik an die Charakteristika realer Datensätze angepasst werden, was häufig nicht vollständig erreicht wird. Dieser Prozess ist arbeitsintensiv und die bestehenden Informationen über Charakteristika realer Daten sind häufig nicht ausreichend [Weiand, 2022, S.6,7].
Manuelle Prüfungen von Datensätzen stellen eine weitere Methode dar. Allerdings unterlaufen auch menschlichen Prüfern Fehler. Dieser Ansatz ist sehr ressourcenintensiv und erfordert viel Zeit. Die Überlegenheit des Menschen gegenüber dem Algorithmus ist nur dann gegeben, wenn ausreichende und zusätzliche Vergleichsdaten verfügbar sind [Doidge et al., 2020, S.8].
Externe Referenzdaten, wie beispielsweise Vergleiche von Sterberaten basierend auf der Verknüpfung von Sterberegistrierungen mit nationalen Zahlen, können ebenso zur Generierung eine Goldstandards herangezogen werden [Harron et al., 2017a, S.8].
Zudem ist es möglich, einen Goldstandard durch die fortlaufende Einbeziehung von Mortalitätsdatenbanken während des Verknüpfungsverfahrens zu erstellen [da Silveira and Artmann, 2009, S.2].
Der Zugriff auf Datensätze, die zur Generierung von Goldstandard-Datensätzen dienen, ist derzeit mit erheblichen Herausforderungen verbunden. Um einen solchen Datensatz zu erhalten, ist in der Regel ein Zugang zu identifizierenden Daten erforderlich, die üblicherweise von vertrauenswürdigen Dritten bereitgestellt werden müssen. Diese Schwierigkeiten, benötigte Daten zu erhalten und einen Goldstandard-Datensatz zu generieren, sind eine bedeutende Hürde für Forscher [Harron et al., 2017b, Harron et al., 2017a, S.1701,1702;9].

3.2. Konsequenzen von Verknüpfungsfehlern

Verknüpfungsfehler in Datensätzen können erhebliche Auswirkungen auf die Ergebnisse und Schlussfolgerungen in Forschungsstudien haben. Die in Kapitel 1.1 aufgeführten Konsequenzen werden in diesem Kapitel detailliert dargestellt [Harron et al., 2017b, S.1700].
Verknüpfungsfehler können zu Verzerrungen (Bias), also zu falschen Einschätzungen des Zusammenhangs zwischen Exposition und Wirkung innerhalb einer spezifischen Population, führen [Kvalsvig et al., 2019, S.12].
Das Ausmaß dieser Auswirkungen variiert je nach Forschungsvorhaben, Art des Fehlers, Anzahl der Fehler, interessierender Variablen sowie der Verteilung der Verknüpfungsfehler auf die interessierenden Variablen [[Doidge and Harron, 2019] zitiert in [Harron et al., 2020, S.219]].
Eine Auswirkung der Homonym- und Synonymfehler können Selektionsverzerrungen sein, sodass die Stichprobe nicht mehr für die Grundgesamtheit repräsentativ ist, da Identitäten wie im Beispiel der Ziwllingpaare fälschlich verknüpft werden oder Identitäten aufgrund von Veränderungen der Daten, wie es beispielsweise bei einer Heirat oder Scheidung der Fall wäre, als nicht zusammengehörig identifiziert werden.
Außerdem können Fehlklassifizierungen entstehen, bei denen die Daten in falsche Kategorien eingeordnet werden. Weitere Auswirkungen sind Messfehler sowie Informationsverzerrungen, die zu falschen Schlüssen aus den Daten führen können [Doidge and Harron, 2019, S.2051].
Die Anzahl der Verknüpfungsfehler ist zweifellos von Bedeutung, jedoch spielt die Verteilung der Fehler auf die interessierende Variable eine entscheidende Rolle, insbesondere im Hinblick auf Selektions- und Informationsverzerrungen. Verknüpfungsfehler, die mit der interessierenden Variable in Verbindung stehen, können zu Fehlklassifizierungen, Messfehlern oder fehlenden Daten führen und damit die Wahrscheinlichkeit von Verzerrungen erhöhen [Doidge and Harron, 2019, S.2053].
Inwieweit Fehler toleriert werden können und wie diese in den Schlussfolgerungen zu bewerten sind, muss grundsätzlich individuell je nach Studienfrage entschieden werden [[Harron et al., 2020, S.219]zitiert in[Doidge and Harron, 2019, S.2053]].

In wissenschaftlichen Untersuchungen spielen falsch negative Ergebnisse (Nachnamensänderung) sowie falsch positive Ergebnisse (Zwillingspaar) eine entscheidende Rolle bei der Bildung einer Stichprobe. Hier können Verknüpfungsfehler zu falschen Ausschlüssen in der Stichprobe führen. Dies hat wiederum Fehlklassifizierungen zur Folge, bei denen Daten fälschlicherweise einer bestimmten Kategorie zugeordnet werden. Außerdem können Messfehler auftreten, die falsche Schlussfolgerungen nach sich ziehen [Doidge and Harron, 2019, S.2058].

Eine Folge dieser Fehlklassifizierungen ist die Informationsverzerrung, die dazu führt, dass die gesammelten Daten nicht mehr die tatsächlichen Gegebenheiten korrekt widerspiegeln. Ebenso kann es zur Selektionsverzerrung kommen, bei der die Stichprobe aufgrund von falsch eingeschlossenen oder ausgeschlossenen Daten nicht mehr repräsentativ für die tatsächliche Population ist [Doidge and Harron, 2019, Kvalsvig et al., 2019, S.2058,9].
Eine Selektionsverzerrung kann ebenfalls entstehen, wenn fehlende Matches von einer oder mehreren interessierenden Variablen abhängen und somit systematisch fehlen [Doidge and Harron, 2019, S.2052].
Es ist wichtig zu beachten, dass falsche Verknüpfungen nur dann zu Fehlklassifizierungen oder Messfehlern führen, wenn die Informationen, die aus der falschen Verknüpfung abgeleitet werden, sich von den Informationen unterscheiden, die aus einer korrekten Verknüpfung gewonnen worden wären [Doidge and Harron, 2019, S.2051].
Wenn die Auswahl der Daten nicht in Beziehung zur interessierenden Variable steht, treten in der Regel keine Verzerrungen auf. Gleiches gilt für Fehlklassifizierungen, die nicht mit der interessierenden Variable in Verbindung stehen [Doidge and Harron, 2019, S.2053].

Fälschliche Einbeziehungen oder Ausschlüsse von Personen innerhalb einer Studienpopulation können außerdem erhebliche Auswirkungen auf die statistische Aussagekraft haben. Dies ist insbesondere der Fall, wenn die Studienpopulation aufgrund von irrtümlich ausgeschlossenen Personen stark reduziert ist [Harron et al., 2020, S.220].

Eine weitere potenzielle Quelle für Selektionsfehler ist die doppelte Zählung einer Person im Falle eines Synonymfehlers (Nachnamensänderung). Auf der anderen Seite tritt eine Unterzählung auf, wenn ein Homonymfehler (Zwillingspaar) vorliegt. Diese Situationen gehen häufig mit Informations- und Auswahlverzerrung einher [Doidge and Harron, 2019, Harron et al., 2020, S.2052,220].

Fehlende Verknüpfungen können zu einer falsch negativen Klassifizierung eines Krankheitsstatus führen. Dies wiederum kann Messfehler und letztendlich eine Unterschätzung der Krankheitsrate zur Folge haben. Zum anderen können aus falschen Verknüpfungen falsch positive Fehlklassifizierungen entstehen und zu einer Überschätzung der Krankheitsrate führen [Doidge and Harron, 2019, Harron et al., 2020, S.2058,220].
In einigen Fällen können daher falsche Übereinstimmungen die Schätzung der Prävalenz beeinflussen, beispielsweise wenn Datensätze von einem Überlebenden und einem Verstorbenen miteinander verknüpft werden [Harron et al., 2017a, S.7].

Das zuvor erwähnte DFG-Projekt in Kapitel 1.1 verdeutlicht die Auswirkungen von Verknüpfungsfehlern auf die Ergebnisse, insbesondere bei der Unterschätzung des Krebsrisikos [Intemann et al., 2023, S.31,89].

Von großer Bedeutung sind Fehler, die systematisch in bestimmten Personengruppen auftreten. Diese systematischen Fehler können Auswirkungen auf die Ergebnisse haben und dazu führen, dass einige Personengruppen über- oder unterrepräsentiert sind, was wiederum das Gesamtergebnis beeinflusst [Harron et al., 2020, S.220].

Es ist jedoch zu beachten, dass die unterschiedlichen Auswirkungen von fehlenden Matches und falschen Matches nicht immer nachteilig sind. In einigen Fällen können sich diese Fehler gegenseitig ausgleichen. Wenn etwa eine Sterblichkeitsrate durch Verknüpfungen bestimmt wird und die Anzahl der falschen und nicht gefundenen Übereinstimmungen sich aufhebt, kann die Rate trotz dieser Fehler korrekt sein [Harron, 2022, Doidge and Harron, 2019, S.2,2051].
Daher ist es von großer Bedeutung, das Ausmaß sowie die Verteilung der Verknüpfungsfehler zwischen den interessierenden Variablen sorgfältig abzuschätzen, um die resultierende Verzwerrung der Ergebnisse hinreichend einschätzen zu können [Harron et al., 2020, S.220].

3.3. Marktanalyse Record Linkage Lösungen

Im Folgenden werden einige frei verfügbare sowie kommerzielle Record Linkage Lösungen hinsichtlich ihrer Verknüpfungsmethode und ihres Einsatzes im Forschungskontext vorgestellt. Der E-PIX, ChoiceMaker, die Mainzelliste, Primat, Link King, FRIL, Febrl, und OpenEMPI gehören zu den kostenfreien Open-Source Record Linkage Lösungen, die im Folgenden vorgestellt werden [Intemann et al., 2023, March et al., 2019, S.167,27].
Kommerzielle Anwendungen sind G-Link, LinkageWiz und DataMatch [March et al., 2019, S.27].
Kommerzielle Lösungen erfordern oft zusätzliche Programmierung, was dazu führen kann, dass die Verknüpfungsumgebung durch die proprietären Verknüpfungsmaschinen eingeschränkt wird [Christen, 2008, S.1].
Im Anhang A der Arbeit ist eine tabellarische Übersicht der Lösungen und ihrer Eigenschaften zu finden.

3.3.1. E-PIX

Der E-PIX wurde durch die Universitätsmedizin Greifswald im Rahmen des GANI-MED-Projekts entwickelt und 2014 erstmals unter Open-Source Lizenz veröffentlicht [Bialke et al., 2015a].
Die Record Linkage Lösung zeichnet sich unter anderem durch die Unterstützung der IHE Profile PIX&PDQ aus und verfügt über eine SOAP sowie eine HL7-FHIR-Schnittstelle [Hampf, 2021, S.7,9].
Der E-PIX schafft Möglichkeiten, Synonymfehler in Personendaten zuverlässig zu erkennen und diese Duplikate auf Basis des Master Patient Index (MPI) zu einer Identität zusammenzuführen, um ein effizientes Identitätenmanagement zu realisieren [Rau et al., 2020, THS, 2022, S.2].
Die Lösung wird kontinuierlich durch die Unabhängige Treuhandstelle Greifswald weiterentwickelt und hat sich insbesondere in nationalen Forschungsvorhaben etabliert [THS, 2023], darunter ausgewählte Krebsregister der Länder, ausgewählte Krankenkassen, Deutsche Zentren für Gesundheit (DZGs) wie das Deutsche Zentrum für Herz-Kreislauf-Forschung (DZHK), die Deutsche Knochenmarkspenderdatei (DKMS), das Berlin Institute of Health (BIH) der Charité Berlin, die größte Langzeistudie Deutschlands „NAKO Gesundheitsstudie“ sowie ein großer Teil der 36 Universitätsstandorte des Netzwerk Universitätsmedizin (NUM) und der Medizininformatik-Initiative (MII).

Innerhalb des E-PIX ist ein Haupt- und Nebenidentitätenkonzept implementiert, das die Verwaltung mehrerer Ausprägungen von IDAT zu einer Person ermöglicht. Jeder Person wird eine Hauptidentität zugeordnet, von der ausgegangen wird, dass diese auf korrekten Daten basiert. Wenn mehrere potenzielle Varianten der IDAT vorliegen, werden alternative Datensätze der Hauptidentität als sogenannte Nebenenidentitäten zugeordnet [Hampf, 2021, S.7,9].
Mit einer grafischen Benutzeroberfläche über den Web-Browser und weiteren Schnittstellen ermöglicht der E-PIX die Registrierung von Patienten. Für jeden registrierten Patient, generiert der E-PIX eine eindeutige Kennung, die als MPI-ID bezeichnet wird und innerhalb der angegebenen Domäne eindeutig ist [Hampf, 2021, Kötzschke, 2015, S.12, 22,32;44-48].
Dabei können zusätzlich zu dem MPI-ID Generator, der bereits im E-PIX implementiert ist, weitere Generatoren hinzugefügt werden [Hampf, 2021, S.31].
Die Konfigurierbarkeit und Erweiterbarkeit des E-PIX ermöglichen eine Anpassung an verschiedene Anwendungsgebiete, unterstützt durch eine umfangreiche Dokumentation [THS, 2023a]. Der Anpassungsprozess kann auf Grundlage einer Standardkonfiguration beginnen, die über das Web-Frontend im XML-Format angepasst werden[Hampf, 2021, Kötzschke, 2015, S.26-28,44-48].
Um verschiedene Dokumentationsformate zu bewältigen, ermöglicht der E-PIX

die Vorbereitung von Datenfeldern für das Record Linkage durch einfache (Ersetzen von Zeichen) und komplexe (Feldweite Operationen) Transformationen [Hampf, 2021, S.34,35].
Bei der einfachen Transformation wird eine spezifische Zeichenkette gesucht und durch eine andere angegebene Zeichenkette ersetzt [Hampf, 2021, S.36].
Bei einer komplexen Transformation kann zwischen verschiedenen Aktionen ausgewählt werden. So gibt es die Möglichkeit alle Kleinbuchstaben durch Großbuchstaben zu ersetzen, Umlaute zu ersetzen oder führende und nachfolgende Leerzeichen zu entfernen [Hampf, 2021, S.36,37].
Der E-PIX verwendet das Blocking als Indexing-Methode und den Fellegi-Sunter-Algorithmus für die Bestimmung der Übereinstimmungswahrscheinlichkeiten der Matching-Variablen. Dabei ermöglicht die Lösung sowohl einen probabilistischen als auch ein deterministsichen Vergleich.[Hampf, 2021, S.37].
Die für das Blocking oder das Matching verwendeten Matching-Variablen können mit Name, Gewicht, Schwellenwert und Ähnlichkeitsmaß definiert werden. Bei der Standardkonfiuguration werden die Matching-Variablen Vorname, Nachname, Geburtsdatum und Geschlecht verwendet, es können jedoch auch weitere Matching-Variablen hinzugefügt werden[Hampf, 2021, Kötzschke, 2015, S.30, 39-42;44-48].
Für das Blocking kann außerdem der Blocking-Mode eingestellt werden, der festlegt, ob es sich um einen Text oder eine Zahl handelt [Hampf, 2021, S.40].
Innerhalb des Abgleichs der Werte einer Matching-Variable werden verschiedene Ähnlichkeitsmaße angeboten:

- Kölner Phonetic Algorithmus
- Deterministischer Algorithmus
- Levenshtein Algorithmus

Eine detaillierte Vorstellung dieser Algorithmen ist in Kapitel 2.2 zu finden.
Weitere zentrale Konfigurationen umfassen Schwellenwerte wie „treshold-possible-match“ und „threshold-automatic-match“ sowie Optionen für das Verhalten des Systems bei Identitätsüberschneidungen.
Wenn ein Identifier mit dem einer Identität identisch ist und mindestens ein Match mit einer andere Person vorliegt, dann stehen verschiedene Optionen zur Verfügung. Darunter die Möglichkeit, wenn mehr als ein Match vorhanden ist, einen Fehler anzuzeigen, dass ein Identifier nur einer Person pro Domäne zugeordnet sein darf und wenn nicht mehr als ein Match vorliegt, die Identität gespeichert und als „Possible Match“ identifiziert wird.
Eine andere Möglichkeit wäre eine Konfiguration, bei der wenn mehr als ein Match vorhanden ist, die Identität gespeichert wird, und als „Possible Match“

gekennzeichnet wird und dies ebenfalls passiert, wenn nicht mehr als ein Match vorliegt [Hampf, 2021, S.27,37,38].
Beim Record Linkage des E-PIX können wie bereits in Kapitel 2.2.4 erläutert, den Datenpaaren abhängig von ihrem Gesamtgewicht und den Schwellenwerten verschiedene Match-Typen zugeordnet werden.

Matching-Typ	**Ereignis**	**Handlung**
Perfect Match	Exakte Übereinstimmung	Automatische Zusammenführung
Automatic Match	„threshold-automatic-match" wurde erreicht/überschritten	Kennzeichnung Haupt- und Nebenidentität
Possible Match	„treshold-possible-match" erreicht/überschritten & „threshold-automatic-match" nicht erreicht/überschritten	Manuelle Prüfung & Generierung einer voläufigen MPI-ID
No Match	„treshold-possible-match" nicht erreicht/überschritten	Generierung einer neuen Person
Multiple Match	Mehrere potenzielle Matches	Generierung einer Liste möglicher Matches

Tabelle 3.1.: Matching-Typen mit zugehörigen Ereignissen und Handlungen im E-PIX [Hampf, 2021, S.15,24](entnommen aus: eigene Aufnahmen)

Eine weitere Konfigurationsmöglichkeit betrifft die Reaktion auf Multiple-Value-Felder, sodass die Herausforderung einer falschen Reihenfolge zweier Zeichenketten adressiert werden kann, indem eine konfigurierbare Gewichtsmenge vom Gesamtgewicht des Datenpaares abgezogen wird, wenn Unstimmigkeiten in der Reihenfolge der Werte vorliegen [Hampf, 2021, S.42].
Der E-PIX bietet außerdem vielfältige Funktionalitäten zur Identifikation und Auflösung potenzieller Synonymfehler. Eine zentrale Komponente ist die manuelle Dublettenauflösung. Für den Abgleich der possible Matches werden die jeweiligen Personendaten übersichtlich gegenübergestellt, sodass manuell entschieden werden kann, ob die Datensätze verknüpft werden, oder nicht [Hampf, 2021, Kötzschke, 2015, S.17,44-48].
Des Weiteren ermöglicht die Record Linkage Lösung den Export und Import sämtlicher Personendaten in einer CSV-Datei, führt ein Protokoll über die ermittelten Match-Typen für jeden Datensatz und stellt domänenübergreifende Statistiken bereit [Hampf, 2021, S.18-21].

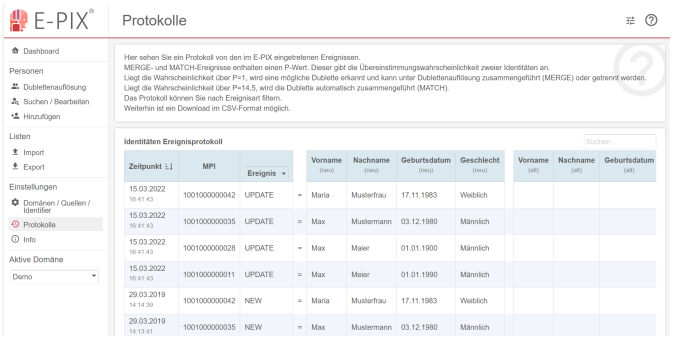

Abbildung 3.1.: Grafische Benutzeroberfläche des E-PIX mit fiktiven Daten (entnommen aus: eigenen Aufnahmen)

3.3.2. ChoiceMaker

ChoiceMaker ist eine Open-Source Record Linkage Lösung, das von ChoiceMaker Technologies mit Sitz in Princeton, New Jersey, USA entwickelt wurde. Diese Java-Anwendung ist über eine Web Service (SOAP)-Schnittstelle und eine Benutzeroberfläche zugänglich und nutzt maschinelles Lernen, um menschliche Matching-Entscheidungen zu imitieren. Die Software besteht aus zwei wesentlichen Komponenten, nämlich dem Analyzer, der die Abgleichlogistik entwickelt und dem Server, der den tatsächlichen Abgleich durchführt. ChoiceMaker basiert auf einem Plugin-Framework, wodurch die Software an spezifische Kundenanforderungen angepasst werden kann [ChoiceMaker, 2023].
Die Lösung unterstützt sowohl einen exakten als auch einen probabilistischen Vergleich. Bei der Bestimmung von Ähnlichkeiten zwischen ausgewählten Merkmalen bietet ChoiceMaker Algorithmen wie Soundex, Editierdistanzen, Jaro-Winkler, New York State Identification and Intelligence System (NYSIIS), Metaphone, Double-Metaphone, Value-Frequency-Weighting und die Levenshtein-Distanz an.
Der ChoiceMaker kommt in verschiedenen Projekten zum Einsatz. Einige Beispiele dafür sind das Children´s Data Network (CDN) an der Universität von Südkalifornien, das Zentrum für städtische Armut und Gemeindeentwicklung der Case Western Reserve University, die Abteilung für Gesundheit und psychische Hygiene in New York City, das Center for Health Record Linkage im Bundesstaat New South Wales in Australien, das Amt für landesweite Gesundheitsplanung und -enticklung im Bundesstaat Kalifornien und die Abteilung für statistische Analyse und Verknüpfung des Gesundheitsministeriums im Bundesstaat Queensland in Australien [ChoiceMaker, 2023].
Vor dem eigentlichen Datenabgleich standardisiert ChoiceMaker die Daten, um einheitliche Grundlagen zu gewährleisten. Der Abgleichprozess erfolgt in zwei Schritten. Zuerst wird das „Blocking“ genutzt, um potenzielle Übereinstimmungen

zu reduzieren. Hierbei kommt der automatisierte Blockierungsalgorithmus von ChoiceMaker zum Einsatz, der dynamisch Gruppen von SELECT-Anweisungen erstellt, die als „Blockierungssätze" bezeichnet werden. Diese Sätze zielen darauf ab, großzügig Daten abzurufen, die mit einem bestimmten Abfragemuster übereinstimmen, ohne jedoch eine vordefinierte Anzahl von Datensätzen zu überschreiten [ChoiceMaker, 2023].
Der maschinelle Lernansatz von ChoiceMaker trainiert die Modelle mit einer Reihe von Datenpaaren, die als „Übereinstimmung", „Unterschied" oder „Unsicher" gekennzeichnet sind. Dabei kommt die Maximum-Entropie-Modellierung (ME) zum Einsatz. Während des Trainingsprozess werden Gewichtungen aller aktiven Hinweise kombiniert, um eine Übereinstimmungswahrscheinlichkeit zu berechnen. Die künstliche Intelligenz von ChoiceMaker analysiert dabei verschiedene Aspekte, wie die Übereinstimmung von Vornamen, phonetische Ähnlichkeiten von Nachnamen und Unterschiede in Geburtsdaten [ChoiceMaker, 2023].
Kunden haben die Möglichkeit, Toleranzen für falsch positive und falsch negative Ergebnisse festzulegen und die Ergebnisse manuell zu überprüfen und anzupassen. ChoiceMaker gibt eine Liste von Datensatz-IDs aus, die die Wahrscheinlichkeit einer Übereinstimmung mit dem Abfragedatensatz anzeigt. Die Entscheidungen werden als „Übereinstimmung" oder „Behalten/Mögliche Übereinstimmung" charakterisiert [ChoiceMaker, 2023].

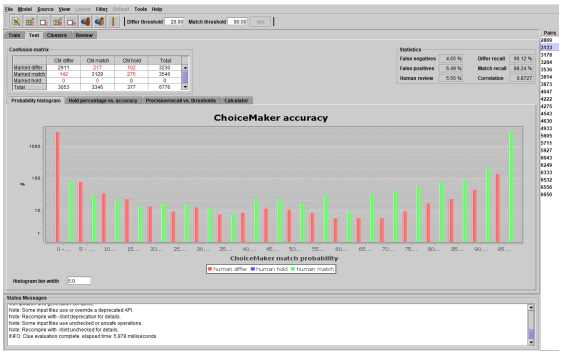

Abbildung 3.2.: Benutzeroberfläche von ChoiceMaker (entnommen aus: [ChoiceMaker, 2023])

3.3.3. Mainzelliste

Die Mainzelliste stellt eine webbasierte Open-Source-Software mit prototypischer grafischer Benutzeroberfläche dar, die von der Mainzelliste Community am Institut für Medizinische Biometrie, Epidemiologie und Informatik (IMBEI) der Universität

Mainz entwickelt wird und die Funktionen der Pseudonymisierung, Pseudonymverwaltung sowie des Record Linkage bereitstellt [Rohde et al., 2021, S.2].
Die Software verfügt über eine REST-basierte Schnittstelle und wird in verschiedenen Kontexten eingesetzt, darunter in nationalen medizinischen Forschungsnetzwerken, zentralisierten Biobanken, Forschungsplattformen, kommerziellen Datenerfassungs- und Analysesuiten, Registersoftwarelösungen sowie Patientenorganisationen in Verbindung mit zugehörigen Krankheitsregistern. Die Mainzelliste ist im Institut für Medizinische Informatik der Universität Münster und im Deutschen Krebsforschungszentrum (DKFZ) in Heidelberg im Einsatz [Rohde et al., 2021, S.2].
Die Record Linkage Lösung verwendet einen gewichtsbasierten und fehlertoleranten Matching-Algorithmus, der Ähnlichkeiten zwischen den Datensätzen identifiziert und Duplikate zusammenführt. Die Daten können in Klartext oder auf Basis kodierter Attributwerten verknüpft werden. Für die Kodierung steht der Bloomfilter zur Verfügung [Rohde et al., 2021, S.2,3].
Vor der eigentlichen Verknüpfung besteht die Möglichkeit zur Datenbereinigung und Transformation. Die Software verwendet verschiedene Ähnlichkeitsmaße für den Vergleich von Daten, darunter die Dice-Ähnlichkeit, die Binäre Ähnlichkeit, die n-Gramm Methode für Textfelder und Wertgleichheit für numerische Variablen. Die Klassifizierung erfolgt anhand von Schwellenwerten.
Bei mehreren potenziellen Übereinstimmungen wird nur der Datensatz mit der höchsten Übereinstimmung berücksichtigt und andere potenzielle Matches werden ignoriert. Die Ähnlichkeit zwischen zwei Datensätzen wird durch die gewichtete Summe aller Matching-Variablen ermittelt. Diese Gewichtung basiert auf der durchnittlichen Häufigkeit und Fehlerquote der Felder.
Die Mainzelliste ermöglicht außerdem eine manuelle Überprüfung und bietet Blocking-Optionen an [Rohde et al., 2021, S.5].
Für die Korrektur doppelter Werte in einem Feld stehen Konfigurationen wie etwa die Aufteilung von Namen mittels Bindestrichen und Leerzeichen zur Verfügung, um die Gesamtähnlichkeit pro Komponente zu bestimmen [Rohde et al., 2021, S.7].
Mit Hilfe einer globalen Kennung (Personenidentifikator (PID)), die jedem Datensatz zugeordnet wird, können potenzielle Duplikate erkannt werden. Liegt ein Duplikat vor, erhält der neu zu registrierende Datensatz die PID des bereits vorhandenen Datensatzes, ansonsten erhält er eine eigene PID und wird in der Datenbank gespeichert [Rohde et al., 2021, S.4].

Abbildung 3.3.: Benutzeroberfläche der Mainzelliste (entnommen aus: [Mainzelliste, 2023])

3.3.4. Primat

Die Open-Source-Toolbox „Primat“ wird von der Universität Leipzig entwickelt und befindet sich derzeit noch in der Entwicklungsphase. Der Fokus von Primat liegt darauf, datenschutzkonformes Matching zu ermöglichen, insbesondere im Kontext des Privacy-Preserving Record Linkage (PPRL). Anders als viele andere Record Linkage Lösungen verfügt Primat derzeit über keine Benutzeroberfläche.
Primat ist flexibel und skalierbar, sodass PPRL-Abläufe an spezifische Projekte angepasst werden können [Franke et al., 2019, S.1826].
Das Linkage-Verfahren nach Fellegi und Sunter wird von Primat unterstützt und es werden vordefinierte Methoden und Parameter, die auf Probedaten basieren, angeboten. Primat ermöglicht darüber hinaus das Blocking, die Kodierung der Daten mittels Bloomfilter oder anderer Methoden wie dem Two-Step Hash Encoding sowie eine optionale Datenbereinigung vor dem Matching.
Die Lösung ist in zwei Hauptbereiche unterteilt: Vorverarbeitung und Linkage. Im Vorverarbeitungsbereich können realistische synthetische Datensätze generiert werden, die zur Feinabstimmung des PPRL-Workflows dienen. In diesem Bereich erfolgen auch die Datenbereinigung und die Kodierung von Attributen zur Wahrung der Privatsphäre [Primat, 2023].
Das Verknüpfungsmodul von Primat enthält vier Hauptkomponenten.
Eine Dienstprogrammkomponente schlägt geeignete Methoden zur Bestimmung von Parametern vor [Franke et al., 2019, S.1828].
Die zweite Komponente „Batch-Verknüpfung” umfasst verschiedene Verknüpfungstechniken sowie Blocking-Methoden wie Standard- und LSH-basiertes Blocking. Für die Klassifizierung stehen binäre Ähnlichkeitsmaße wie die Jaccard-Ähnlichkeit, die Dice-Ähnlichkeit und die Hamming-Ähnlichkeit zur Verfügung. Bei mehreren potenziellen Übereinstimmungen wird der Datensatz mit der höchsten Übereinstimmung verwendet [Franke et al., 2019, S.1828].
Die dritte Komponente ist die inkrementelle Verknüpfung, die eine Nutzung von Da-

tenbanken zur Speicherung und Abfrage früherer Matching-Ergebnisse ermöglicht. Die vierte Bewertungskomponente erlaubt es Nutzern Laufzeit, Reduktionsverhältnis und Blockgröße zu bewerten und auf unvorteilhafte Parameter oder Methoden hinzuweisen [Franke et al., 2019, S.1829].
Primat strebt außerdem an, verschiedene Nachbereitungsmethoden, wie die symmetrische beste Übereinstimmung, das stabile Matching und die Maximum-Gewichtsanpassung bereitzustellen, um die Qualität der Ergebnisse zu verbessern. Da sich Primat aktuell noch in der Entwicklungsphase befindet, sind einige Funktionen noch in der Umsetzung. Dazu gehören Aspekte wie Datenkorruption, Intra-Source-Datensatzverknüpfung, Benutzeroberfläche, Vorverarbeitungsvorlage, Privater Schemaabgleich und die inkrementelle Verknüpfung [Primat, 2023].

3.3.5. Link King

Die Lösung „Link King", entwickelt von der Abteilung für Alkohol- und Drogenmissbrauch des Bundesstaates Washington, ist eine Statistical Analysis System (SAS) Anwendung, die sich auf die Verknüpfung und Entkoppelung von Verwaltungsdatensätzen spezialisiert. Die Anwendung selbst ist kostenfrei, erfordert jedoch eine Lizenz für Base SAS, die 2000$ kostet [Dusetzina SB, 2014, S.49].
Link King bietet die Möglichkeit sowohl probabilistisches als auch deterministisches Record Linkage umzusetzen. Die deterministischen Protokolle wurden in verschiedenen Evaluations- und Forschungsprojekten der Abteilung für Alkohol- und Drogenmissbrauch erfolgreich angewendet.
Die Lösung ermöglicht interaktive und Stapelverarbeitungsmodi und setzt künstliche Intelligenz ein, um Verknüpfungsprotokolle zu überprüfen und den Prozess der Datenaufbereitung und Verknüpfung zu optimieren [Campbell, 2005, LinkKing, 2023, S.2,5]
Link King verlangt die Pflichtvariablen Vorname, Nachname und Geburtsdatum oder Sozialversicherungsnummer. Die Anwendung ermöglicht über eine Benutzeroberfläche die manuelle Prüfung unsicherer Verknüpfungen und die Konfiguration von Gewichten. Es werden Wahrscheinlichkeitsalgorithmen verwendet, die ursprünglich für das Datenbankenprojekt der Substance Abuse and Mental Health Administration von MEDSTAT entwickelt wurden [Campbell, 2005, S.1-3].
Link King verfügt über eine umfassende Funktionalität zur Qualitätssicherung der Daten. Vor dem Abgleich können Formatstrukturen angegeben werden, beispielsweise für die Kodierung des Geschlechts. Ein weiteres Beispiel ist die Möglichkeit, fehlende Geschlechtsdaten automatisiert zu ergänzen. Hierfür werden Sterbeurkunden aus den Jahren 1980 bis 2003 im US-Bundesstaat Washington als Thesauren verwendet. Zusätzlich enthält Link King Listen mit gültigen Werten für Geburtsdaten und Sozialversicherungsnummern, um fehlerhafte Werte zu identifizieren[Campbell, 2005, S.2].

Methoden wie die Soundex Funktion von SAS, String-Vergleichsfunktionen und die phonetische Äquivalenzfunktion des New York State Intelligence Information System werden genutzt, um falsch geschriebene Namen zu identifizieren.
Der Link King kann außerdem Spitznamen identifizieren. Dies geschieht über einen weiteren Thesaurus, welcher Spitznamen enthält und das Hinzufügen von weiteren Spitznamen erlaubt [Campbell, 2005, S.2].
Link King ordnet, basierend auf dem Verknüpfungsprotokoll und dem Unsicherheitsgrad, verknüpfte Datensätze in 11 Kategorien ein. Der Benutzer hat die Möglichkeit, mittels des festgestellten Fehlergrads und Zufallsstichproben, in jeder Kategorie zu entscheiden ob diese eingeschlossen oder ausgeschlossen werden sollen [Campbell, 2005, S.3].

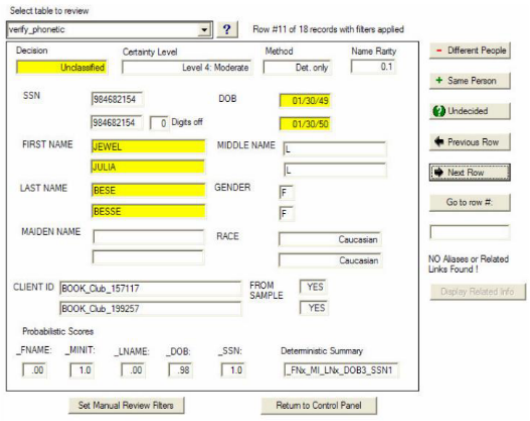

Abbildung 3.4.: Benutzeroberfläche von LinkKing (entnommen aus: [Campbell, 2005])

3.3.6. FRIL

„FRIL" ist eine Open-Source-Lösung, die im Rahmen einer Zusammenarbeit zwischen der Emory University und den Centers for Disease Control and Prevention in Atlanta, Georgia, entwickelt wurde [FRIL, 2023].
FRIL findet Anwendung bei Forschungsgruppen des National Center on Addiction and Substance Abuse, der Harvard Business School und des National Institute of Health (NIH). Die Lösung ermöglicht sowohl einen probabilistischen als auch einen deterministischen Abgleich [Jurczyk et al., 2008, FRIL, 2023, S.442].
Bevor der Datenabgleich stattfindet, können verschiedene Transformationen auf die Datensätze angewendet werden. Darunter die Zusammenführung zweier Attribute in der Datenquelle zu einem Attribut [FRIL, 2023].
Eine Benutzeroberfläche ermöglicht die Auswahl einer Suchmethode, eines

Ähnlichkeitsmaßes und eines Entscheidungsmoduls für den Datensatzabgleich [Jurczyk et al., 2008, S.441].
Die Suchmethode bestimmt, welche Datenpaare zwischen den Quellen verglichen werden sollen. FRIL bietet verschiedene Suchmethoden wie den Nested Loop Join und die Sortierte Nachbarschaftsmethode.
Auch für Attribute mit mehrfachen Werten wie doppelte Vornamen ermöglicht FRIL die Aufteilung oder Normalisierung der Attributwerte während des Verknüpfungsprozess [Jurczyk et al., 2008, S.441].
Die Lösung unterstützt verschiedene Ähnlichkeitsmaße wie Editierdistanzen, Soundex, n-Gramm und Gleichheit. Der Benutzer kann für jede einzelne Variable ein passendes Ähnlichkeitsmaß auswählen. FRIL ermöglicht außerdem die Konfiguration von Distanzfunktionen mittels Fuzzy Logik, sodass der Benutzer Schwellenwerte für Übereinstimmung und Nicht-Übereinstimmung festlegen kann. Automatische Gewichtungsvorschläge können manuell angepasst werden [Jurczyk et al., 2008, S.441].
Für deterministisches Record Linkage kann die Gleichheit als Distanzfunktion festgelegt werden, wodurch nur exakt übereinstimmende Variablen als zusammengehörig definiert werden.
Zudem verfügt FRIL über grafische Tools zur Analyse, Validierung und Zusammenfassung der Ergebnisse. Die Ergebnisse werden in die Kategorien „Match“ , „Possible Match“ und „No Match“ sortiert.
Aktuell wird daran gearbeitet, FRIL um mehrere automatisierte Tools, darunter Techniken des maschinellen Lernens, zu erweitern, um Werte für bestimmte Parameter vorzuschlagen [Jurczyk et al., 2008, S.441,442,444].

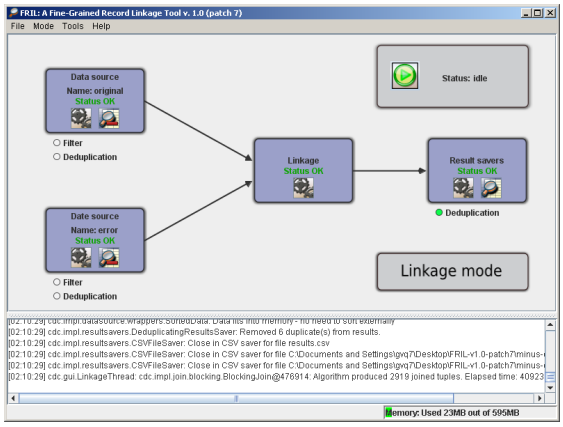

Abbildung 3.5.: Benutzeroberfläche von FRIL (entnommen aus: [FRIL, 2023])

3.3.7. Febrl

„Febrl" ist eine kostenlose Open-Source Software, die von der Australian National University Data Mining Group in Canberra entwickelt wurde und im Gesundheitswesen Anwendung findet. Febrl eignet sich besonders für kleine bis mittlerer experimentelle Verknüpfungen und Deduplizierungen[Christen, 2008, S.3,8].
Die Benutzeroberfläche von Febrl ermöglicht sowohl deterministisches als auch probabilistisches Record Linkage und bietet Funktionen wie die Datenbereinigung, Verknüpfung und Deduplizierung von Datensätzen sowie die Möglichkeit des manuellen Abgleichs von Datenpaaren. Über die grafische Benutzeroberfläche können Parameter für die Datensatzverknüpfung konfiguriert werden [Christen, 2008, S.3,6,7].
Die integrierte Datenbereinigungsfunktion bietet Komponentenstandardisierer für Namen, Adressen, Daten und Telefonnummern. Außerdem erlaubt die Software die Definition von Blockierungsschlüsseln und stellt sieben verschiedene Indizierungsmethoden bereit, wie n-Gramm Index, Fuzzy-Blocking, CanopyIndex, StringMapIndex und SuffixArrayIndex. Für den Vergleich von Datensätzen bietet Febrl 26 verschiedene Vergleichsfunktionen, darunter String-Vergleichsfunktionen und Vergleiche für numerische Werte. Die Anpassung von Übereinstimmungs- und Nicht-Übereinstimmungsgewichten ist ebenfalls möglich [Christen, 2008, S.4,5].
Febrl erlaubt die Auswahl verschiedener Klassifikationstechniken, darunter den Fellegi und Sunter Klassifikator und den „OptimalThreshol" Klassifikator, bei dem ein optimaler Schwellenwert auf Basis des summierten Gewichtsvektors berechnet werden kann. Die angebotenen Identifikatoren „Kmeans" und „FarthestFirst" basieren auf Clustering-Ansätzen und teilen die Gewichtsvektoren in einen Match- und Non-Match-Cluster ein. Weiterer mögliche Klassifikatoren sind „SuppVecMachine" und „Two Step" [Christen, 2008, S.6].
Außerdem besteht die Möglichkeit die Ausgabe der Ergebnisse zu konfigurieren.

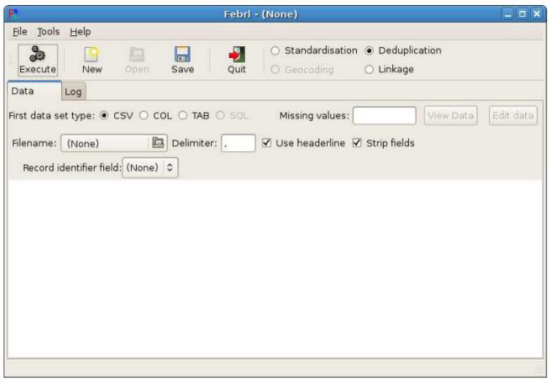

Abbildung 3.6.: Benutzeroberfläche von Febrl (entnommen aus: [Christen, 2008])

3.3.8. OpenEMPI

„OpenEMPI" repräsentiert eine Open-Source-Implementierung eines Enterprise Master Patient (EMPI), zeichnet sich durch die Unterstützung des deterministischen als auch des probabilistischen Record Linkage aus und verfügt über eine REST-basierte Schnittstelle. Die Software wurde von der SYSNET International entwickelt [OpenEMPI, 2023a].
Die Flexibilität von OpenEMPI zeigt sich in der Möglichkeit, Konfigurationsparameter auf die spezifischen Anforderungen einzelner Projekte anzupassen. Durch eine erweiterbare Architektur können verschiedene Matching-Algorithmen in die Software integriert werden. Auch eine Datenbereinigung vor dem tatsächlichen Vergleich ist möglich [OpenEMPI, 2023a].
Die Software verfügt über verschiedene Blockierungs- und Matching-Algorithmen. Im Kontext der Blockierungskonfiguration können Matching-Felder spezifiziert werden. Als Suchmethode kann neben dem Blocking auch die sortierte Nachbarschaftsmethode verwendet werden.
Für das Matching selbst können über die Benutzeroberfläche Schwellenwerte, Ähnlichkeitsmaße, Matching-Variablen und Gewichte konfiguriert werden. Eine manuelle Verknüpfung ist ebenfalls möglich [OpenEMPI, 2023b].
Für die Bestimmung der Ähnlichkeit zweier Zeichenketten kann ein deterministischer Algorithmus verwendet werden sowie probabilistische Maße, wie beispielsweise die Levenshtein-Distanz oder der Jaro-Winkler Algorithmus. Diese können implementiert werden, ohne dass Änderungen an dem gesamten System erfolgen müssen.
Die neuste Version (OpenEMPI 4.3.0) wurde im September 2023 veröffentlicht und bietet einen auf maschinellen Lernen basierenden Algorithmus. Dieser Algorithmus soll die beste Ähnlichkeitsmetrik und den optimalen Schwellenwert für einen Datensatz bestimmen. Zudem wurden verschiedene Datenstandardisierungsfunktionen hinzugefügt, um fehlende oder fehlerhafte Daten in den Datensätzen zu identifizieren und zu korrigieren. Die Software eröffnet eine Auswahl verschiedener Bibliotheken und präsentiert auf der Startseite Diagramme, die Informationen über die Verknüpfung bereitstellen. Dies ermöglicht Benutzern eine frühzeitige Erkennung von Problemen [OpenEMPI, 2023b].

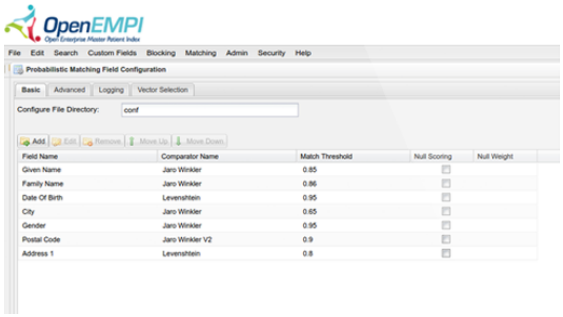

Abbildung 3.7.: Benutzeroberfläche von OpenEMPI (entnommen aus: [OpenEMPI, 2023a])

3.3.9. G-Link

„G-Link" von Statistics Canada, ist eine Software, die für 12.500 kanadische Dollar erhältlich ist und sich vor allem im Bereich der Gesundheitswissenschaften etabliert hat. Die Anpassungsfähigkeit von G-Link zielt darauf ab, Datensätzen ohne eindeutigen Identifikator zuverlässig zu verknüpfen [Fair, 2004, S.37,43].
G-Link unterstützt deterministische Algorithmen, den NYSIIS-Algorithmus und Tippfehler-Vereinbarungen. Der probabilistische Verknüpfungsprozess der Lösung gliedert sich in drei Phasen, die über eine Benutzeroberfläche konfiguriert werden können. In der ersten Phase, der Suchphase, werden Benutzerdatensätze hochgeladen. Durch das Blocking entsteht eine Zufallsstichprobe nicht verknüpfter Paare, die zur Berechnung nicht verknüpfter Ergebnisgewichtungen dient. In dieser Phase werden Regeln für Übereinstimmungen festgelegt und ein feldweiser Vergleich durchgeführt. Analysetabellen können dabei helfen, relevante Datenmerkmale zu identifizieren, die bei der Auswahl von Blockierungsvariablen hilfreich sind. Die Klassifizierung der Paare basiert auf Schwellenwerten [Bellow et al., 2016, S.3229].
In der zweiten Phase, der Entscheidungsphase, werden Verknüpfungsregeln auf die potenziellen Paare angewendet. Der Benutzer hat die Möglichkeit Quotenverhältnisse für verknüpfte Paare sowie Schwellenwerte anzupassen und Wahrscheinlichkeiten neu festzulegen. Diese Anpassungen ermöglichen eine Neu-Klassifizierung der Paare. Die Untersuchung von Stichproben und die Nutzung verfügbarer grafischer Anzeigen können bei der Wahl der Schwellenwerte unterstützen [Fair, 2004, Bellow et al., 2016, S.43,44;3229].
In der Gruppenphase ordnet G-Link die Datensätze entsprechend ihres Verknüpfungsstatus. Das Programm erkennt Konflikte, wenn ein Datensatz in einer Datei mit mehreren Datensätzen in einer anderen Datei verknüpft ist. Die Konfliktlösungen sind sowohl automatisch als auch manuell möglich [Bellow et al., 2016, S.3229].

3.3.10. LinkageWiz

„LinkageWiz", entwickelt von LinkageWiz Software in Australien, wird in verschiedenen Bereichen wie der Medizin und der Sozialforschung eingesetzt. Die Software ist in mehreren Organisationen in den USA, Kanada, Großbritannien, Australien und Frankreich im Einsatz und für ein Limit von 50.000 Datensätzen für 199$ sowie für ein unbegrenztes Datensatzlimit für 2.999$ erhältlich [LinkageWiz, 2023].
Als Ähnlichkeitsmaße werden phonetische Algorithmen wie NYSIIS und Soundex sowie String-Vergleichsfunktionen angeboten. Neben dem probabilistischen Datenabgleich unterstützt LinkageWiz auch einen deterministischen Abgleich. Die Lösung verwendet Identifikatoren wie Name, Geburtsdatum, Geschlecht, Adresse, Sozialversicherungsnummer und Vornamen. Darüber hinaus kann der Benutzer die Gewichtung der verwendeten Variablen anpassen. Die Konfigurationen können über eine Benutzeroberfläche vorgenommen werden [LinkageWiz, 2023].
LinkageWiz bietet ein Modul für die Datenbereinigung und -transformation, um Daten in ein einheitliches Format zu überführen und Abkürzungen zu identifizieren. Mit Hilfe einer Standardisierung können Datensätze beispielsweise in Groß- und Kleinschreibung umgewandelt, Städte und Bundesländer auf Abkürzungen reduziert, Leerzeichen und weitere Satzzeichen entfernt oder nur Ziffern für ein Feld zulässig gemacht werden [LinkageWiz, 2023].
Fehlende Angaben zum Geschlecht können auf Basis einer umfangreichen Bibliothek mit mehr als 30.000 Einträgen von Vornamen abgeleitet werden.
LinkageWiz enthält einen Assistenten, der die Adressbereinigungen erleichtert und verfügt über eine umfassende Bibliothek von Thesauren für verschiedene Datenelemente wie Straßentypen, Ordnungszahlen, Grundstückstypen, Suffixe, Postfächer, Vororte und Städte. Eine Geokodierungs-Engine ermöglicht die Geokodierung von Adressdaten.
Darüber hinaus ist ein Import der Datensätze und ein Export der Ergebnisse in Tabellenform möglich [LinkageWiz, 2023].

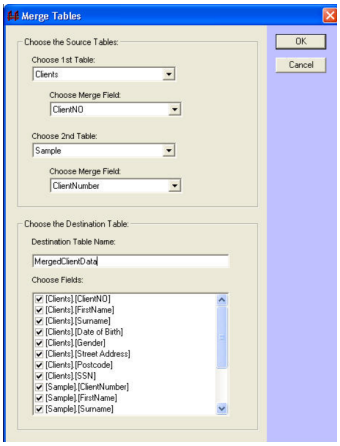

Abbildung 3.8.: Benutzeroberfläche von LinkageWiz (entnommen aus: [LinkageWiz, 2023])

3.3.11. DataMatch Enterprise

„DataMatch Enterprise" ist eine Lösung von DataLadder, die über eine Benutzeroberfläche eine Vielzahl von Anpassungsmöglichkeiten für Datenabgleiche bietet. Diese Lösung wird im Gesundheitswesen und in Branchen wie Finanzierung, Versicherung, Verkauf oder Marketing eingesetzt und verfügt über eine RESTful-API.[DataMatch, 2023].
DataMatch ermöglicht einen probabilistischen und einen deterministischen Datenabgleich und bietet Konfigurationen im Prozess der Datenbereinigung an. Im Rahmen der Datenbereinigung können leere Werte, Leerzeichen, bestimmte Buchstaben oder Zahlen entfernt oder ersetzt werden. Zudem ermöglicht DataMatch die Analyse von Datenfelder, um zwei Informationen innerhalb eines Datenfelds zu trennen und auf zwei Spalten aufzuteilen [DataMatch, 2023].
Die Lösung verfügt über eine Musterbibliothek, um gültige und ungültige Werte zu identifizieren. Die Konfigurationen für den Abgleich ermöglichen die Erstellung mehrerer Übereinstimmungsdefinitionen, die logische Ausdrücke wie UND/ODER-Kriterien verwenden, um Datensätze abzugleichen. Jede Definition kann mehrere Kriterien beinhalten, die auf spezifische Anforderungen zugeschnitten sind. Außerdem können die Gewichtungen für Matching-Felder angepasst werden und es stehen verschiedene passende Algorithmen zur Auswahl, sei es numerisch, phonetisch oder domänenspezifisch, um den Abgleich entsprechend der Daten zu gestalten.
Außerdem werden verschiedene Ähnlichkeitsmaße angeboten: die Levenshtein-Distanz, Damerau-Levenshtein-Abstand, Jaro-Winkler Algorithmus,Tastaturabstand, Kullback-Leibler-Abstand, Jaccard-Ähnlichkeit, Metaphon

3, Name Variante, Silbenausrichtung und Akronym [DataMatch, 2023].
DataMatch liefert Übereinstimmungsergebnisse in Form von Scores, die den Grad des Übereinstimmungsvertrauens in Prozent anzeigen.
Feineinstellungen von Schwellenwerten und Konfidenzniveaus dienen dazu, Fehlinterpretationen zu reduzieren. Die Software ermöglicht zudem Matching-Algorithmen mit unterschiedlichen Schwellenwerten erneut auszuführen, um den optimalen Wert zu wählen, der die geringste Anzahl von falsch positiven und falsch negativen Ergebnissen gewährleistet. Während des prozess ermöglicht eine Live-Vorschau den Anwendern die Auswirkungen der Datenbereinigung und des Abgleichs in Echtzeit zu überprüfen, um gewünschte Ergebnisse sicherzustellen. Die Plattform bietet zudem eine automatische Planung für den Datenabgleich [DataMatch, 2023].

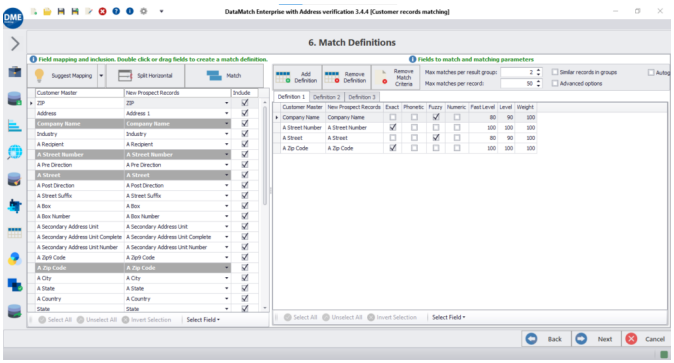

Abbildung 3.9.: Benutzeroberfläche von DataMatch (entnommen aus: [DataMatch, 2023])

3.3.12. Fazit der präsentierten Lösungen

Die präsentierten Record Linkage Lösungen bieten vielfältige Ansätze und Funktionen für den Datenabgleich. Sie weisen jedoch Unterschiede in der Anwendung von Algorithmen, in den Konfigurationsoptionen und der Art der Datenbereinigung und -transformation auf.
Einige Lösungen eignen sich ausschließlich für das probabilistische Record Linkage, während die Mehrzahl sowohl das probabilistische als und das deterministische Modell unterstützt.
In Bezug auf die Datenbereinigung und -transformation bieten mehrere Lösungen ähnliche Möglichkeiten, wie das Entfernen von Satzzeichen oder Leerzeichen. ChoiceMaker, Link King und DataMatch integrieren künstliche Intelligenz, um den Verknüpfungsprozess zu optimieren. LinkKing nutzt außerdem Thesauren zur automatischen Anpassung des Geschlechts oder zur Identifizierung ungültiger Ge-

burtsdaten.
Die vorgestellten Lösungen präsentieren Benutzeroberflächen, die die Anpassung des Verknüpfungsprozess erleichtern. Ein Großteil der Lösungen beschränkt sich jedoch auf ein einfaches Design.
Viele Lösungen ermöglichen den manuellen Abgleich von potenziellen Matches. In Bezug auf die Suchmethode setzen die meisten Lösungen auf den Blocking-Prozess, während einige Record Linkage Lösungen wie LinkageWiz, die Nachbarschaftsmethode verwenden.
Unterschiede zeigen sich auch in der Konfiguration des probabilistischen Vergleichs. So bieten einige Lösungen wie beispielsweise FRIL einen automatischen Gewichtsvorschlag. Auch in der Auswahl der Ähnlichkeitsmaße wie der Levenshtein-Distanz oder der n-Gramm-Methode unterscheiden sich die Anwendungen. Einige Lösungen unterstützen mehrere Maße, während sich andere Anwendungen auf ein kleineres Angebot beschränken. Da die Art der verwendeten Algorithmen die Vergleichbarkeit von Datensätzen stark beeinflusst und zu unterschiedlichen Ergebnissen führen kann, ist eine angemessene Auswahl der Ähnlichkeitsmaße von entscheidender Bedeutung für die Güte der Analyse. Selbst im Falle einer identischen Konfiguration des Abgleichs und eines identischen Ähnlichkeitsmaßes können die Lösungen unter anderem aufgrund von Unterschieden in der Datenbereinigung, der Datentransformation und der Handhabung mehrerer möglicher Matches zu unterschiedlichen Ergebnisse kommen.
Zusammenfassend lässt sich sagen, dass sich alle präsentierten kommerziellen und Open-Source-Lösungen in ihren Konfigurationsmöglichkeiten ähneln. Dennoch heben sich die Record Linkage Lösungen voneinander ab, da unterschiedliche Methoden für die Datenbereinigung, Transformation, Konfiguration sowie Auswahl der Ähnlichkeitsmaße verwendet werden und einige Lösungen eine künstliche Intelligenz integriert haben, um den Verknüpfungsprozess zu optimieren.

4. Systematische Entwicklung eines Fehlerminimierungskonzepts

4.1. Durchführung

Diese Untersuchung zielt darauf ab, Verknüpfungsergebnisse der Record Linkage Lösungen E-PIX und FRIL anhand eines Goldstandard-Datensatzes zu evaluieren. Der Goldstandard-Datensatz dient als Referenz zum Vergleich der Effektivität beider Record Linkage Lösungen hinsichtlich der Minimierung von Homonym- und Synonymfehlern. Basierend auf den entwickelten Konfigurationen soll neben einem Vergleich der Verknüpfungsqualität eine Empfehlung für eine geeignete Konfiguration im Kontext des Record Linkage für Register abgeleitet werden.
Im Rahmen der Fehlerquellenanalyse wurden, mit Blick auf die Datenqualität, Verknüpfungskriterien entwickelt, die innerhalb des Matching-prozess der jeweiligen Lösungen berücksichtigt werden sollten, um die Qualität der Verknüpfung zu steigern:

1. Tippfehler
2. Zahlendreher
3. Zwillingspaare
4. Adressänderung
5. Multiple-Value-Felder
6. Nachnamensänderung
7. Unterschiedliche Verwendung von Diakritika
8. Geschwister mit gleichem Geburtstag

Die Evaluierung der Record Linkage Lösungen konzentriert sich insbesondere auf die Erfassung von Homonym- und Synonymfehlern vor dem Hintergrund der aufgeführten Kriterien. Dies ermöglicht eine Identifikation potenzieller Optimierungsbereiche im Record Linkage Prozess.

Der vorliegende Datensatz basiert auf einer Teilmenge von IDAT der Treuhandstelle des Klinischen Krebsregisters Mecklenburg-Vorpommern. Der Datensatz

umfasst 7510 Einträge, die bereits durch Melderegister-Vergleiche geprüft wurden und daher als verlässliche Grundlage dienen können.
Im Kontext des Abgleichs mit einem Melderegisters erhielt das Krebsregister Informationen aus dem zentralen Identifikationsregister (ZIR), die mit den im Krebsregister vorhandenen IDAT abgeglichen wurden. Hierfür wurde eine Abfrage aller im E-PIX des Krebsregisters erfassten Personen beim Melderegister durchgeführt. Die Abfrage beim ZIR erfolgte über eine eMRA-Schnittstelle. Hierbei kam ein speziell vom Krebsregister selbst entwickeltes Tool zum Einsatz, das die individuelle Abfrage jeder Person über die eMRA-Schnittstelle ermöglichte. Nach den Einzelabfragen jeder Person wurden die Resultate aufbereitet und anschließend in den E-PIX importiert.
Der vorliegende Datensatz besteht aus einmalig registrierten Identitäten ohne Duplikate sowie manuell zusammengeführte Identitäten, die mit mindestens einer anderen Identität verknüpft wurden. Die Daten setzen sich aus drei verschiedenen Ereignissen zusammen.

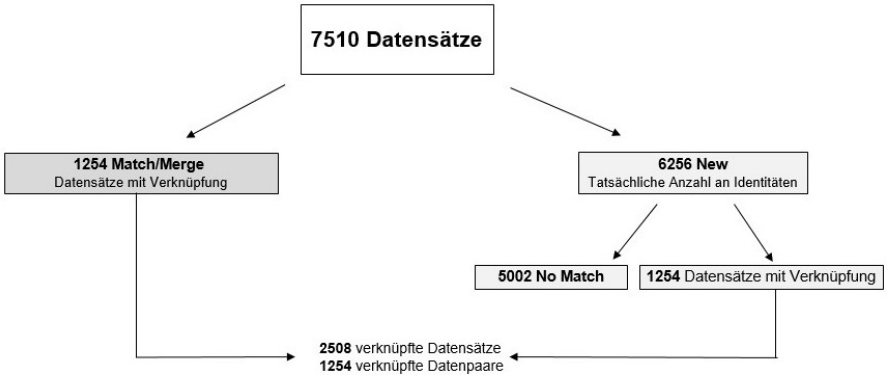

Abbildung 4.1.: Zusammensetzung des Goldstandard-Datensatzes (entnommen aus: eigene Aufnahmen)

Der vorliegende Datensatz beinhaltet verschiedene Matching-Variablen wie Name, Nachname, Geburtsdatum, Geburtsort, Postleitzahl, Geschlecht, Sterbedatum, Sterbeort und Geburtsname. In einigen Werten treten Tippfehler sowie Multiple-Value-Felder auf, die von Record Linkage Lösungen berücksichtigt werden sollten.

In einem ersten Schritt wurden die Record Linkage Lösungen hinsichtlich ihrer Konfigurationsmöglichkeiten verglichen, die in den nachfolgenden Tests verwendet wurden.
Als Zweites wurden Konfigurationen für die Record Linkage Lösungen E-PIX und FRIL entwickelt, die in Anhang B und C dokumentiert wurden. Hierfür wurden die Algorithmen, Gewichte und Schwellwerte an den vorliegenden Datensatz angepasst, um möglichst optimale Ergebnisse zu erzielen. Die Anpassung der Gewichte und

Schwellwerte erfolgte auf Basis eines Abgleichs der ermittelten Fehler mit dem Referenzdatensatz. Die Datentransformation wurde dabei nicht geändert.
Es wurden gezielte Modifikationen vorgenommen, um die Anzahl von Synonymfehlern zu minimieren, wobei jedoch stets darauf geachtet wurde, dass die Anzahl von Homonymfehlern nicht signifikant zunimmt. Denn im Kontext von Registern ist eine genaue Identifikation von Patienten und die konkrete Zuordnung von spezifischen Krankheiten von entscheidender Bedeutung.
Für beide Tools wurden Konfigurationen mit verschiedenen Algorithmen getestet, und analysiert welche Verknüpfungsergebnisse mit den Ergebnissen des Goldstandards übereinstimmen.
Um nicht ausschließlich die Kriterien „Tippfehler“ und „Multiple-Value-Felder“ zu evaluieren, wurden in einem dritten Schritt zusätzliche Datensätze hinzugefügt, um sämtliche in Aufzählung 4.1 aufgeführten Aspekte abzudecken. Dabei wurden die Konfigurationen der Record Linkage Lösungen erneut an den erweiterten Datensatz angepasst und die auftretenden Fehler analysiert. Die Anzahl der Einträge des Datensatzes stieg mit der Erweiterung von 7510 auf 7527.
Da das Klinische Krebsregister Mecklenburg-Vorpommern, analog zu weiteren Registern, die Matching-Variablen Vorname, Nachname, Geschlecht, Geburtsdatum und Postleitzahl verwendet, wurden mit beiden Record Linkage Lösungen weitere Konfigurationen entwickelt und evaluiert, die auf den genannten Matching-Variablen basieren. Es ergaben sich somit folgende Test-Durchläufe für jede Record Linkage Lösung:

Tests	Datensatz	Matching-Variablen	Algorithmen
Test-Durchlauf 1	Goldstandard	Vorname, Nachname, Geschlecht, Geburtsdatum	Levenshtein/Kölner Phonetik/Deterministisch
Test-Durchlauf 2	Goldstandard	Vorname, Nachname, Geschlecht, Geburtsdatum, Postleitzahl	Levenshtein/Kölner Phonetik/Deterministisch
Test-Durchlauf 3	Goldstandard + weitere Testfälle	Vorname, Nachname, Geschlecht, Geburtsdatum	Levenshtein/Kölner Phonetik/Deterministisch
Test-Durchlauf 4	Goldstandard + weitere Testfälle	Vorname, Nachname, Geschlecht, Geburtsdatum, Postleitzahl	Levenshtein/Kölner Phonetik/Deterministisch

Tabelle 4.1.: Durchführung der Tests für jede Record Linkage Lösung (entnommen aus: eigene Aufnahmen)

Im vierten Schritt wurden die wahren positiven Fälle (TP), die wahren negativen Fälle (TN), die falsch positiven Fälle (FP) sowie die falsch negativen Fälle (FN) gezählt, um die Metriken Recall und Precision zu berechnen.
Der fünfte Schritt umfasst den Vergleich der Ergebnisse beider Lösungen. In diesem Schritt werden Schlussfolgerungen zu möglichen Optimierungsbereichen gezogen und die Verknüpfungsqualität der Lösungen gegenübergestellt.

4.2. Ergebnisse

4.2.1. Gegenüberstellung Konfiguration

Zunächst erfolgt ein Vergleich der Konfigurationsmöglichkeiten von E-PIX und FRIL. FRIL bietet zwei Modi: den „Linkage-Mode“ für die Verknüpfung von Daten aus verschiedenen Quellen und den „Deduplication-Mode“ für die Prüfung von Duplikaten in einem Datensatz. Für den Vergleich mit dem E-PIX wurde der Deduplication-Mode verwendet, um den Goldstandard-Datensatz des Krebsregisters auf Duplikate zu prüfen.

Wie in Kapitel 2.2.1 beschrieben, stellt die Wahl der Matching-Variablen den ersten Schritt im Record Linkage Prozess dar. Sowohl der E-PIX als auch FRIL erlauben die freie Wahl der Variablen aus dem vorliegenden Datensatz.

Als zweiter Schritt erfolgt die Konfiguration der Datenbereinigung und Datentransformation, wobei in beiden Lösungen Konfigurationsoptionen vorhanden sind, die sich jedoch in ihrer Umsetzung unterscheiden.
In beiden Lösungen besteht jedoch die Möglichkeit, Regeln für die Datenformate innerhalb einer Matching-Variable festzulegen.
Im E-PIX können frei definierbare Zeichen durch andere ersetzt werden, während die Transformationen in FRIL als Regeln für die Verwendung der Datenfelder gelten. So werden in FRIL die Felder eines Datensatzes nur dann verwendet, wenn sie die frei definierten Regeln wie etwa „Der Wert x muss mit Wert y übereinstimmen“ oder „Der Wert x muss größer sein als Wert y“ erfüllen.
Ergänzend ermöglicht der E-PIX die Anwendung vordefinierter komplexer Transformationen, wie beispielsweise die Entfernung aller führenden und nachfolgenden Leerzeichen, die FRIL innerhalb des Deduplizierungs Modus nicht anbietet.

Im dritten Schritt erfolgt die Konfiguration des Indexing-prozess, wobei beide Record Linkage Lösungen das Blocking als Indexing-Verfahren anbieten.
Innerhalb der XML-Datei des E-PIX kann für jede Matching-Variable spezifiziert werden, ob diese als Blocking-Variable verwendet werden soll. Es besteht außerdem die Möglichkeit, mehrere Variablen für diesen Prozess auszuwählen. Zusätzlich kann für jede Blocking-Variable angegeben werden, ob die Daten als Zahl oder als Wort vorliegen.
Im Vergleich dazu erlaubt FRIL nur die Auswahl einer einzigen Variable, die für das Blocking verwendet werden soll. Für diese Blocking-Variable kann festgelegt werden, ob der Soundex-Code oder das Präfix abgeglichen werden soll.
Die Konfiguration eines Schwellenwertes für die Blocking-Variable, die mit dem E-PIX möglich ist, kann in FRIL nicht erfolgen. Hier kann lediglich die Länge des

Soundex-Codes und des Präfix angegeben werden.

Im vierten Schritt des Matching-prozess erfolgt die Durchführung des Datenabgleichs, wobei sowohl im E-PIX als auch in FRIL Konfigurationsanpassungen erforderlich sind. Wie zuvor erwähnt, ist die Auswahl der Matching-Variablen in beiden Tools frei konfigurierbar. Für jede Variable können unterschiedliche Algorithmen festgelegt werden, wobei sowohl die Verwendung desselben Algorithmus für alle Variablen als auch die Auswahl verschiedener Algorithmen möglich ist.
E-PIX und FRIL unterscheiden sich in ihrem Angebot an Algorithmen für den Datenabgleich. FRIL stellt acht verschiedene Algorithmen zur Verfügung, wobei sich drei auf spezifische Datenformate wie Geburtsdatum oder Adresse beziehen und daher nicht universell auf alle Matching-Variablen anwendbar sind, während der E-PIX drei Algorithmen implementiert hat.
Beide Lösungen bieten die Levenshtein-Distanz, Soundex und einen deterministischen Vergleich an. Zusätzlich hat FRIL das Jaro-Winkler-Maß und die n-Gramm-Methode integriert. Eine Erläuterung der Funktionsweisen dieser Algorithmen ist in Kapitel 2.2.3 zu finden.
Die Konfiguration der Schwellenwerte für die einzelnen Matching-Variablen spielt im Rahmen des Vergleichs der Datensätze eine entscheidende Rolle. In diesem Konfigurationschritt unterscheiden sich die beiden Lösungen.
Im E-PIX werden ein oberer und ein unterer Schwellenwert im Bereich von 0 bis 1 festgelegt. Ein oberer Schwellenwert von 1 bedeutet, dass die Werte identisch sein müssen, um als übereinstimmend identifiziert zu werden.
Im Gegensatz dazu setzt FRIL zwei Schwellenwerte zwischen 0 und 1, wobei 1 bedeutet, dass der größtmögliche Unterschied in den zu vergleichenden Werten akzeptiert wird, während 0.1 als oberer Schwellenwert nur sehr geringe Unterschiede zulassen würde, um eine Übereinstimmung zu identifizieren.
Als abschließender Konfigurationsschritt der Matching-Variablen kann das Gewicht konfiguriert werden. Dabei unterscheiden sich die Herangehensweisen der beiden Lösungen erneut. Während im E-PIX die freie Wahl des Gewichts möglich ist, muss bei FRIL die Gesamtgewichtsmenge aller Matching-Variablen stets 100 ergeben.

Im finalen Schritt des Matching-prozess erfolgt die Klassifizierung der Datenpaare. Im E-PIX kann im Rahmen der Klassifizierung ein oberer und ein unterer Schwellenwert festgelegt werden. Datenpaare, deren Gesamtgewicht innerhalb dieses Intervalls liegt, werden als „possible Match“ betrachtet und einer manuellen Prüfung unterzogen.
Im Deduplizierungsmodus von FRIL kann lediglich ein Schwellenwert im Bereich von 0 bis 100 konfiguriert werden. Sobald dieser Schwellenwert erreicht oder überschritten wird, identifiziert FRIL ein Datenpaar als zusammengehörig. Es

ist zu beachten, dass in diesem Modus die Möglichkeit der Klassifizierung als „possible Match“ fehlt. Dies könnte dazu führen, dass in bestimmten Szenarien Synonymfehler übersehen werden, da sie als Nicht-Übereinstimmung klassifiziert werden und keine Möglichkeit besteht, einen manuellen Abgleich durchzuführen, da dieser im Deduplizierungsmodus von FRIL nicht implementiert ist.

In der Berechnung des Gesamtgewichts eines Datenpaares weisen die Lösungen minimale Unterschiede auf.
In beiden Ansätzen wird das Gewicht jeder Matching-Variable, wie in Abschnitt 2.2.4 erläutert, mit dem Ähnlichkeitswert der vorliegenden Daten multipliziert. Dabei fließen auch die Schwellenwerte der Matching-Variablen in die Berechnung des Gesamtgewichts ein.
Im E-PIX erfolgt nach der Multiplikation eine Differenzierung der Produkte, die sich aus dem Gewicht und dem Ähnlichkeitsmaß der jeweiligen Matching-Variable ergeben haben. Anschließend werden die übereinstimmenden Produkte durch die nicht übereinstimmenden Produkte dividiert.
Im Gegensatz dazu fließen in FRIL die Schwellenwerte in die Berechnung des Ähnlichkeitswertes, der vom Algorithmus ausgegeben wird, mit ein.
Dadurch entsteht ein Score für das abgeglichene Datenpaar. Dieser Score wird anschließend mit dem jeweiligen Gewicht der Variable multipliziert und die Produkte werden zu einer Summe addiert, die das Gesamtgewicht des Datenpaares ausweist. Die Berechnung des Scores für ein Datenpaar kann unter Verwendung der Levenshtein-Distanz in FRIL wie folgt dargestellt werden [Jurczyk, 2009, S.30]:

$$score(strA,strB) = \begin{cases} 0, if\, e(strA,strB) > d * \max(length(strA),length(strB)) \\ 1, if\, e(strA,strB) < a * \max(length(strA),length(strB)) \\ \dfrac{d * \max(length(strA),length(strB)) - e(strA,strB)}{(d-a) * \max(length(strA),length(strB))}, otherwise \end{cases}$$

Abbildung 4.2.: Formel für die Berechnung des Scores eines zu vergleichenden Datenpaares in FRIL unter Verwendung der Levenshtein-Distanz (entnommen aus [Jurczyk, 2009])

strA	strB	e(strA,strB)	Approve level (a)	Disapprove level (d)	Score
"A"	"A"	0	0.1	0.3	1
"A"	"B"	1	0.1	0.3	0
"ADAM"	"ADAMS"	1	0.1	0.3	0.5
"JACOB DOBBS"	"JAKOB HOBBS"	2	0.1	0.3	0.59
"JASPER CISNEROS"	"ADEN CISNEROS"	4	0.1	0.3	0.17
"BREANNA ROBISON"	"BRENNA ROBINSON"	2	0.1	0.3	0.83

Tabelle 4.2.: Beispiele der Berechnung des Scores eines zu vergleichenden Datenpaares in FRIL unter Verwendung der Levenshtein-Distanz (reproduziert von [Jurczyk, 2009])

Zur Verdeutlichung der Formel wird das Datenpaar in der grau markierten Spalte der Tabelle 4.2 betrachtet.
In Spalte e(strA,strB) wird angegeben, dass nur eine Operation erforderlich ist, um das erste Wort in das zweite Wort zu überführen. Vorab wurden für die Werte dieser Matching-Variable spezifische Regelungen getroffen: Bei einem Schwellenwert von 0,1 wird die Änderung genehmigt, während ein Schwellenwert von 0,3 oder höher eine Ablehnung signalisiert.
Gemäß der festgelegten Regeln in Abbildung 4.2 wird der Score auf null gesetzt, wenn die Anzahl der Veränderungen größer ist als das Produkt aus dem ablehnenden Schwellenwert und der maximalen Länge der beiden Wörter. In diesem Fall liegt die Anzahl der Veränderungen bei 1 und das Produkt aus dem ablehnenden Schwellenwert und der maximalen Länge beträgt 0,3. Da die Anzahl der Veränderungen in diesem Fall größer ist, ergibt sich ein Score von null für dieses Datenpaar. Dieser Score wird anschließend mit dem Gewicht der Matching-Variable multipliziert[Jurczyk, 2009, S.30].
Die Berechnungsmethoden des Gesamtgewichts weisen in den Record Linkage Lösungen teilweise Unterschiede auf, jedoch berücksichtigen beide Ansätze alle relevanten Faktoren für die Ermittlung des Gesamtgewichts.

Zusammenfassend lässt sich feststellen, dass der E-PIX bei einigen Konfigurationsmöglichkeiten detailliertere Optionen bereitstellt als FRIL es im Modus der Duplizierung tut. Beide Record Linkage Lösungen bieten jedoch die für den Record Linkage Prozess bedeutenden Konfigurationsmöglichkeiten, wie eine Datentransformation, das Indexing und die Konfiguration der Matching-Variablen, an.

4.2.2. Tests E-PIX

4.2.2.1. Test-Durchlauf 1

E-PIX Durchlauf 1	**Algorithmus**	**TP**	**TN**	**FP**	**FN**	**Recall**	**Precision**	**F-Score**
Konfiguration 1	Levenshtein-Distanz	1254	6254	2	0	1	0,9984	0,9992
Konfiguration 2	Kölner Phonetik	1252	6254	2	2	0,9984	0,9984	0,9984
Konfiguration 3	Deterministisch	1251	6254	2	3	0,9976	0,9984	0,9980

Tabelle 4.3.: Testergebnisse des E-PIX für Test-Durchlauf 1 (entnommen aus: eigene Aufnahmen)

Die Anwendung der Standard-Konfiguration im E-PIX auf den Goldstandard-Datensatz diente als Ausgangspunkt für die Entwicklung weiterer Konfigurationen,

wie sie in Tabelle 4.3 dargestellt sind. Die Ergebnisse der Standard-Konfiguration zeigten eine geringe Anzahl von Synonymfehlern, während gleichzeitig 17 Homonymfehler identifiziert wurden, bei denen Identitäten irrtümlich verknüpft wurden.
Die Fehleranalyse der Standard-Konfiguration ergab, dass unterschiedliche Personen zusammengeführt wurden, obwohl ihre Geburtsdaten nicht übereinstimmen. Außerdem konnten Multiple-Value-Felder mit der Standardkonfiguration nicht berücksichtigt werden. Dies ist dann problematisch, wenn in vier Datensätzen identische Werte für beispielsweise Matching-Variablen wie Nachname, Geburtsdatum und Geschlecht vorliegen, jedoch bei zwei dieser Datensätze „Anna Lena" als Vorname angegeben ist und in den zwei weiteren Datensätzen nur „Anna" als Vorname vorliegt. In solchen Fällen wurden fälschlicherweise alle vier Datensätze als eine Person identifiziert. Dieser spezifische Fehler hat maßgeblich zu den falsch positiven Ergebnissen beigetragen.

Auf Basis der Evaluation der Standard-Konfiguration wurde eine Konfiguration erstellt, welche den Levenshtein-Algorithmus verwendet. Dabei wurden die Gewichte und Schwellenwerte an die Charakteristik des vorliegenden Datensatz angepasst. Wie bereits im vorangegangenen Abschnitt 3.3.1 erläutert, bietet der E-PIX die Möglichkeit über eine spezielle Funktion, eine konfigurierbare Gewichtsmenge vom Gesamtgewicht eines Datenpaares abzuziehen, wenn bestimmte Teile eines Multiple-Value Feldes nicht übereinstimmen. Durch eine entsprechende Erhöhung der konfigurierbaren Gewichtsmenge konnten alle Synonymfehler identifiziert und die Anzahl der Homonymfehler signifikant auf zwei reduziert werden.
Verbleibende Homonymfehler ergeben sich aus Fällen, bei denen vier identische Datensätze vorliegen und fälschlicherweise zu einer Identität durch den E-PIX verknüpft wurden, wobei die Reaktion des E-PIX grundsätzlich als korrekt betrachtet werden kann. Zusätzliche Matching-Variablen wären erforderlich, um diese Fälle als zwei Identitäten zu identifizieren.
Insgesamt erfüllt der E-PIX unter Verwendung der Levenshtein-Distanz, nach Anpassungen der Standard-Konfiguration an den Datensatz sowie der Anpassung des Gewichts eines Datenpaares im Falle von Multiple-Value-Feldern, die Kriterien „Tippfehler" und „Multiple-Value-Felder", die in dem vorliegenden Datensatz enthalten sind.

Unter Anwendung der Kölner Phonetik ergeben sich jeweils zwei Homonym- und Synonymfehler. Die beiden Homonymfehler korrelieren mit denen, die unter Verwendung der Levenshtein-Distanz entstanden sind. Die Anzahl der Synonymfehler ist im Vergleich zu den Ergebnissen der Konfiguration mit der

Levenshtein-Distanz von null auf zwei gestiegen. Diese Synonymfehler resultieren aus Tippfehlern in den Vornamen der Identitäten. Aufgrund dieser Schreibfehler weist der Algorithmus den Wörtern unterschiedliche Codes zu, was zu einer Nichtübereinstimmung führt. Infolgedessen werden die betroffenen identitäten als nicht zusammengehörig klassifiziert. Die Überprüfung der Verknüpfungskriterien ergab, dass das Kriterium Multiple-Value-Felder zu berücksichtigen, erfüllt ist. Im Gegensatz dazu konnte das Kriterium „Tippfehler" nicht erfüllt werden.

Der deterministische Abgleich zeigt vergleichbare Ergebnisse mit denen der Kölner Phonetik und weist zusätzlich einen weiteren Synonymfehler auf.
Die Synonymfehler basieren ebenfalls auf Tippfehlern in den Matching-Variablen „Vorname" und „Nachname". Aufgrund der Unterschiede in den Werten der Matching-Variablen werden die Datensätze nicht als Übereinstimmung identifiziert. Analog zu den Ergebnissen unter Verwendung der Kölner Phonetik konnte mit dieser Konfiguration korrekt auf Multiple-Value-Feld reagiert werden, jedoch konnte das Kriterium „Tippfehler" nicht erfüllt werden.

Der E-PIX erzielte unter Verwendung der Levenshtein-Distanz in Verbindung mit den Pflicht-Matching-Variablen die höchste Verknüpfungsqualität. Mit Hauptaugenmerk auf der Vermeidung von falsch negativen Ergebnissen, erzielte die erste Konfiguration einen Recall von 1 und eine Precision von 0,99. Im Vergleich dazu wies der deterministische Vergleich die geringste Verknüpfungsqualität auf.

4.2.2.2. Test-Durchlauf 2

Im Rahmen des Test-Durchlaufs 2 erfolgte eine Erweiterung der zuvor entwickelten Konfigurationen, die ausschließlich die Pflichtvariablen des E-PIX verwendeten. In diesem Test-Durchlauf wurde die Postleitzahl als zusätzliche Matching-Variable hinzugefügt.

E-PIX Durchlauf 1	**Algorithmus**	**TP**	**TN**	**FP**	**FN**	**Recall**	**Precision**	**F-Score**
Konfiguration 1	Levenshtein-Distanz	1254	6254	2	0	1	0,9984	0,9992
Konfiguration 2	Kölner Phonetik	1252	6254	2	2	0,9984	0,9984	0,9984
Konfiguration 3	Deterministisch	1251	6254	2	3	0,9976	0,9984	0,9980

Tabelle 4.4.: Testergebnisse des E-PIX für Test-Durchlauf 2 (entnommen aus: eigene Aufnahmen)

Unter Verwendung der Levenshtein-Distanz konnten in diesem Test-Durchlauf erfolgreich alle Homonymfehler aufgelöst werden, die im ersten Testdurchlauf auftraten. Die Hinzunahme der Postleitzahl als zusätzliche Information führte zu

einem Unterschied in ansonsten identischen Datensätzen und ermöglichte somit die Behebung von Fehlern, bei denen der E-PIX fälschlicherweise Identitäten als zusammengehörig identifizierte. Die Anzahl der Synonymfehler ist im Vergleich zum Test-Durchlauf 1 um einen Fehler angestiegen. Dieser Fehler bezieht sich auf zwei Datensätze, bei denen alle Angaben, bis auf die Postleitzahl, identisch sind und die Datensätze tatsächlich dieselbe Person repräsentieren.
Trotz dieses Anstiegs erfüllte der E-PIX, unter Verwendung der Levenshtein-Distanz und der Pflichtvariablen mit zusätzlicher Berücksichtigung der Postleitzahl, die Kriterien „Tippfehler“ und „Multiple-Value-Felder“ erneut. Das Kriterium „Adressänderung” konnte durch die Hinzunahme der Postleitzahl ebenfalls getestet werden. Dieses wurde unter Verwendung der Konfiguration 4 nicht vollständig berücksichtig.

Die Anwendung des Algorithmus „Kölner Phonetik“ zeigte ähnliche Fehler wie Konfiguration 2 im ersten Test-Durchlauf. Die Fehleranalyse ergab, dass das Kriterium „Tippfehler“ unter Verwendung dieses Algorithmus weiterhin nicht erfüllt werden konnte. Das Kriterium Multiple-Value-Felder zu berücksichtigen bleibt unverändert erfüllt. Die Hinzunahme der Postleitzahl hat in diesem Fall daher keine Auswirkungen auf die Ergebnisse. Das Kriterium „Adressänderung” konnte unter Verwendung der Kölner Phonetik erfüllt werden.

Die dritte Konfiguration nutzte den deterministischen Abgleich und erzielte im Vergleich zum ersten Test minimal verbesserte Ergebnisse. Unter Verwendung der Pflichtvariablen blieben im ersten Test-Durchlauf zwei Homonymfehler und drei Synonymfehler bestehen, während in Durchlauf 2 keine Homonymfehler und vier Synonymfehler vorliegen. Diese Synonymfehler sind auf Tippfehler in den Matching-Variablen „Vorname“ und „Nachname“ sowie auf eine Adressänderung zurückzuführen, wie es bei einem deterministischen Abgleich zu erwarten ist. Jedoch konnte auch mit dieser Konfiguration das Kriterium Multiple-Value-Felder zu berücksichtigen erfüllt werden.

Zusammenfassend lässt sich festhalten, dass die Hinzunahme der Matching-Variable „Postleitzahl“ zuvor vorhandene Homonymfehler erfolgreich aufgelöst hat. Allerdings führte die Einbeziehung der weiteren Variable dazu, dass die Anzahl der Synonymfehler minimal anstieg. Die Reduzierung der Homonymfehler hatte in diesem Fall keinen Einfluss auf die Erhöhung der Synonymfehler. Die Messwerte für Recall und Precision bewegen sich analog zu den Testergebnissen des ersten Durchlaufs im Bereich von 0,99 und 1, was auf eine hohe Verknüpfungsqualität hinweist. Auch hier zeigt sich die höchste Verknüpfungsqualität unter Anwendung der Levenshtein-Distanz.

4.2.2.3. Test-Durchlauf 3

Im Rahmen der Testdurchläufe 3 und 4 wurde der Referenzdatensatz durch die Integration weiterer Testdatensätze von 7510 auf 7527 Einträge erweitert, um sämtliche in Aufzählung 4.1 aufgeführten Aspekte abzudecken. In diesem Test-Durchlauf 3 wurden ausschließlich die Matching-Variablen Namen, Nachname, Geschlecht und Geburtsdatum verwendet.

E-PIX Durchlauf 3	**Algorithmus**	**TP**	**TN**	**FP**	**FN**	**Recall**	**Precision**	**F-Score**
Konfiguration 7	Levenshtein-Distanz	1258	6262	4	3	0,9976	0,9968	0,9972
Konfiguration 8	Kölner Phonetik	1256	6264	2	5	0,9960	0,9984	0,9972
Konfiguration 9	Deterministisch	1251	6264	2	10	0,9921	0,9984	0,9952

Tabelle 4.5.: Testergebnisse des E-PIX für Test-Durchlauf 3 (entnommen aus: eigene Aufnahmen)

Die Konfiguration 7 setzte die Levenshtein-Distanz als Matching-Algorithmus ein. Mit der Integration zusätzlicher Datensätze stieg wie zu erwarten die Anzahl der Homonym- und Synonymfehler im Vergleich zu den Ergebnissen des ersten Testdurchlaufs, in dem der Datensatz nicht erweitert wurde und ebenfalls Pflichtvariablen eingesetzt wurden.
Die Homonymfehler resultierten aus Fällen, in denen alle Werte der Matching-Variablen übereinstimmten, jedoch die Postleitzahl unterschiedlich war. Da die Postleitzahl in diesem Durchlauf nicht verwendet wurde, reagiert der E-PIX analog zu Test 1 grundsätzlich korrekt, indem er diese Fälle zu einer Identität zusammenführt, obwohl es sich um zwei unterschiedliche Identitäten handelt.
Die Synonymfehler entstanden durch Nachnamensänderungen, die der Algorithmus nicht korrekt identifizierte. Dagegen konnte der E-PIX Tippfehler, Zahlendreher, Zwillingspaare, Multiple-Value-Felder, die unterschiedliche Verwendung von Diakritika und Geschwister, die am gleichen Tag Geburtstag haben, korrekt identifizieren. Die Nachnamensänderung konnte in diesem Fall nicht vollständig berücksichtigt werden, da eine korrekte Identifizierung die Anzahl der Homonymfehler stark erhöht hätte. Das Kriterium „Adressänderung“ wurde in diesem Test-Durchlauf nicht geprüft, da die Postleitzahl nicht als Matching-Variable verwendet wurde.

Die Konfiguration 8 nutzte die Kölner Phonetik und wies, ähnlich wie unter Verwendung der Levenshtein-Distanz, eine Zunahme von Verknüpfungsfehlern auf. Die Anzahl der Homonymfehler blieb gleich, während die Zahl der Synonymfehler von zwei auf fünf anstieg. Die Homonymfehler bestanden wie in Test-Durchlauf 1 aus identischen Datensätzen, bei denen lediglich ein Unterschied in der Postleitzahl vorlag. Die Synonymfehler resultierten aus Tippfehlern und geänderten Nachnamen. Alle weiteren Fälle, die die Verknüpfungskriterien abbilden,

konnten korrekt identifiziert werden.

Konfiguration 9 verwendete den deterministischen Abgleich und zeigte ebenfalls ein Zunahme der Verknüpfungsfehler im Vergleich zu Durchlauf 1. Während die Zahl der Homonymfehler konstant bei zwei blieb, stieg die Anzahl der Synonymfehler von drei auf zehn an. Die Homonymfehler sind mit denen in Test 1 identisch. Die Synonymfehler bestehen aus Tippfehlern, Zahlendrehern, Nachnamensänderungen und einer unterschiedlichen Verwendung von Diakritika. Das Zwillingspaar, die Geschwister, die am gleichen Tag Geburtstag haben, die Multiple-Value-Felder und die Adressänderungen konnten jedoch erfolgreich berücksichtigt werden.
Die Synonymfehler verdeutlichen, dass die Kriterien, die zur Verbesserung der Verknüpfungsqualität eingehalten werden sollten, mit diesem Algorithmus zum Großteil nicht erfüllt werden können.

Trotz der schwer zu identifizierenden Fälle wiesen alle Konfigurationen eine hohe Verknüpfungsqualität auf. Diese Qualität wurde erneut anhand der Kennzahlen Recall und Precision bewertet und erreichte in allen Fällen einen Wert von ungefähr 0,99. Die Levenshtein-Distanz erzielte weiterhin die besten Ergebnisse und wies vor allem die geringste Anzahl an Synonymfehlern auf, gefolgt von der Kölner Phonetik und dem deterministischen Abgleich.

4.2.2.4. Test-Durchlauf 4

Im Zuge des Testdurchlaufs 4 wurde ebenfalls der erweiterte Datensatz als Referenz verwendet. Im Unterschied zum vorherigen Durchlauf 3 wurde in diesem Test die Matching-Variable „Postleitzahl" erneut hinzugezogen.

E-PIX ***Durchlauf 4***	**Algorithmus**	**TP**	**TN**	**FP**	**FN**	**Recall**	**Precision**	**F-Score**
Konfiguration 10	Levenshtein-Distanz	1260	6263	3	1	0,9992	0,9976	0,9984
Konfiguration 11	Kölner Phonetik	1255	6264	2	6	0,9952	0,9984	0,9968
Konfiguration 12	Deterministisch	1250	6266	0	11	0,9913	1	0,9956

Tabelle 4.6.: Testergebnisse des E-PIX für Test-Durchlauf 4 (entnommen aus: eigene Aufnahmen)

Die Konfiguration 10 nutzte die Levenshtein-Distanz und wies aufgrund der Hinzunahme weiterer Fälle erwartungsgemäß zu einem minimalen Anstieg der Verknüpfungsfehler, im Vergleich zu Test-Durchlauf 2, in dem ebenfalls die Postleitzahl als weitere Matching-Variable verwendet wurde.
Die Integration der Postleitzahl reduzierte die Zahl der Homonymfehler von vier in Testdurchlauf 3 auf drei in Testdurchlauf 4, während die Anzahl der

Synonymfehler von drei auf einen zurückging. Die Homonymfehler bestanden erneut aus Fällen mit identischen Werten, jedoch unterschiedlichen Postleitzahlen, die in Test-Durchlauf 2 durch die Hinzunahme der Postleitzahl identifiziert werden konnten. Da sich die Anzahl der Synonymfehler bei einer korrekten Identifizierung dieser Homonymfehler erhöht hätte und die Priorität auf der Vermeidung der Synonymfehler liegt, wurden diese Homonymfehler im vorliegenden Test akzeptiert. Der Synonymfehler resultierte aus einer Nachnamensänderung, wobei Nachnamensänderungen bei anderen Datensätzen korrekt identifiziert wurden, während einer dieser Fälle weiterhin bestehen blieb. Dies ist auf die Ähnlichkeit der Nachnamen zurückzuführen. Eine nähere Betrachtung des Algorithmus legt nahe, dass bei einer größeren Ähnlichkeit in der Schreibweise der Nachnamen der übrig gebliebenen Synonymfehler identifiziert worden wären.
Insgesamnt konnte unter Verwendung dieser Konfiguration die Kriterien Tippfehler, Zahlendreher, Zwillingspaar, Multiple-Value-Felder, die unterschiedliche Verwendung von Diakritika und Geschwister, die am gleichen Tag Geburtstag haben, korrekt identifiziert werden. Hingegen konnten Adressänderungen und Nachnamensänderungen nicht vollständig berücksichtigt werden.

Die Konfiguration 11 setzte die Kölner Phonetik ein und zeigte ebenfalls die zu erwartende Zunahme der Verknüpfungsfehler im Vergleich zu Durchlauf 1.1, in dem der Datensatz nicht um weitere Fälle ergänzt wurde. Die Anzahl der Homonymfehler blieb konstant, während die Zahl der Synonymfehler von zwei auf sechs anstieg. Die Integration der Postleitzahl führte, im Vergleich zu Test-Durchlauf 3, zu einer Erhöhung der Anzahl von Synonymfehlern von fünf auf sechs. Die Homonymfehler waren erneut durch Fälle charakterisiert, in denen alle Werte bis auf die Postleitzahl identisch waren, während die Synonymfehler durch Adressänderungen, Nachnamensänderungen und Tippfehlern bedingt waren. Alle anderen Fälle, die die Matching Kriterien abbildeten, wurden erfolgreich identifiziert.

Die Konfiguration 12 wies unter Verwendung des deterministischen Abgleichs die gleiche Anzahl an Verknüpfungsfehlern auf wie in Test-Durchlauf 3, in dem die Postleitzahl nicht berücksichtigt wurde. Die Verknüpfungsfehler beruhen auf den Kriterien Adressänderung, Tippfehler, Nachnamensänderung, Zahlendreher und der unterschiedlichen Verwendung von Diakritika.
Wie bereits in den vorherigen Testdurchläufen erzielte die Levenshtein-Distanz erneut die höchste Verknüpfungsqualität und zeigte die geringste Anzahl an Synonymfehler, während der deterministische Abgleich in allen Tests die geringste Verknüpfungsqualität und die höchste Anzahl an Synonymfehler aufwies.

4.2.3. Tests FRIL

Im Rahmen der Evaluation der Record Linkage Lösung FRIL fand die Klassifikation ebenso wie unter Verwendung des E-PIX mit dem Fellegi-Sunter Algorithmus statt. Außerdem wurden für den Abgleich der Datensätze die gleichen Algorithmen wie in den Tests mit dem E-PIX verwendet, darunter die Levenshtein-Distanz, Soundex und ein deterministischer Abgleich. Die entwickelten Konfigurationen sind in Anhang C zu finden.
Innerhalb der Tests mit dem E-PIX wurde die Kölner Phonetik als eine Ausprägung des Algorithmus Soundex verwendet, da diese Ausprägung auf deutsche Worte spezialisiert ist. Obwohl FRIL keine spezifischen Informationen darüber bereitstellt, ob eine spezifische Ausprägung des Soundex verwendet wird, deuten die Verknüpfungsergebnisse darauf hin, dass dieser auch mit deutschen Worten umgehen kann.

4.2.3.1. Test-Durchlauf 1

Für den ersten Test wurden die Konfigurationen an den Goldstandard-Datensatz des Krebsregisters angepasst, wobei die Matching-Variablen Vorname, Nachname, Geburtsdatum und Geschlecht verwendet wurden.

FRIL Durchlauf 1	**Algorithmus**	**TP**	**TN**	**FP**	**FN**	**Recall**	**Precision**	**F-Score**
Konfiguration 1	Levenshtein-Distanz	1254	6254	2	0	1	0,9984	0,9992
Konfiguration 2	Soundex	1253	6250	6	1	0,9992	0,9952	0,9972
Konfiguration 3	Deterministisch	1251	6254	2	3	0,9976	0,9984	0,9980

Tabelle 4.7.: Testergebnisse von FRIL für Test-Durchlauf 1 (entnommen aus: eigene Aufnahmen)

Die erste Konfiguration nutzte die Levenshtein-Distanz als Algorithmus und konnte alle Synonymfehler (Multiple-Value-Felder und Tippfehler) identifizieren. Die zwei entstandenen Homonymfehler bilden sich aus Fällen, in denen jeweils vier Datensätze exakt identische Werte aufweisen. In diesem Szenario verhält sich FRIL grundsätzlich korrekt.
Für eine korrekte Einordnung dieser Fälle wäre eine Integration weiterer Matching-Variablen erforderlich.

Die zweite Konfiguration verwendete Soundex und konnte im Gegensatz zu Konfiguration 1 einen Synonymfehler nicht identifizieren.
Dieser Fehler resultierte aus einem Schreibfehler in einem Nachnamen. Die Funktionsweise des Algorithmus lässt darauf schließen, dass die Wörter aufgrund des Tippfehlers nicht ähnlich klingen und als Folge unterschiedliche Codes für die

Wörter generiert wurden, die eine Nicht-Übereinstimmung signalisieren.
Zwei der Homonymfehler korrelierten mit denen, die unter Verwendung der Levenshtein-Distanz entstanden sind, während zusätzliche Homonymfehler aufgrund von Multiple-Value-Feldern aufgetreten sind.
Die Kriterien „Tippfehler" und „Multiple-Value-Felder" konnten unter Verwendung von Soundex nur teilweise berücksichtigt werden.

Die dritte Konfiguration nutzte einen deterministischen Vergleich, wodurch die Anzahl der Homonymfehler im Vergleich zu Konfiguration 2 abnahm, die Anzahl der Synonymfehler jedoch zunahm.
Die zwei Homonymfehler bestehen erneut aus identischen Datensätzen, während die vier Synonymfehler aus Schreibfehlern entstanden. Diese Fehler resultieren aus der Funktionsweise des Algorithmus, bei der Datenfelder als nicht zusammengehörig identifiziert werden, sofern sie nicht vollständig identisch sind.
In diesem Kontext wurde das Kriterium, Multiple-Value-Felder zu berücksichtigen zwar erfüllt, jedoch erfolgte eine unzureichende Reaktion auf Tippfehler.

In diesem ersten Test-Durchlauf zeigte die Konfiguration, die den Levenshtein-Algorithmus implementierte, mit einem Recall von 1 und einer Precision von 0,9984 die höchste Verknüpfungsqualität.
Der deterministische Abgleich erzielte in diesem Durchlauf mit Blick auf die Synonymfehler die geringste Verknüpfungsqualität. Es ist jedoch zu beachten, dass im Gesamtkontext der Verknüpfungsqualität Soundex minimal schlechtere Ergebnisse erzielte als der deterministische Abgleich.

4.2.3.2. Test-Durchlauf 2

In diesem Test-Durchlauf wurden die Konfigurationen um die Matching-Variable „Postleitzahl" erweitert. Diese Ergänzung ermöglichte es, neben den Multiple-Value-Feldern und den Tippfehlern auch das Kriterium der Adressänderung einzubeziehen.

FRIL ***Durchlauf 2***	**Algorithmus**	**TP**	**TN**	**FP**	**FN**	**Recall**	**Precision**	**F-Score**
Konfiguration 4	Levenshtein-Distanz	1254	6254	2	0	1	0,9984	0,9992
Konfiguration 5	Soundex	1253	6250	6	1	0,9992	0,9952	0,9972
Konfiguration 6	Deterministisch	1250	6256	0	4	0,9968	1	0,9984

Tabelle 4.8.: Testergebnisse von FRIL für Test-Durchlauf 2 (entnommen aus: eigene Aufnahmen)

Unter Verwendung der Levenshtein-Distanz in Konfiguration 4 ergab sich durch die Hinzunahme der Postleitzahl als Matching-Variable keine Veränderung der

Verknüpfungsfehler.
Die Homonymfehler der Konfiguration 1 konnten trotz der Hinzunahme der Postleitzahl nicht vermieden werden. Dies liegt an der Zunahme der Synonymfehler, zu der eine korrekte Identifizierung der beiden Homonymfehler geführt hätte.

Die Ergebnisse der fünften Konfiguration ähneln den Ergebnissen der zweiten Konfiguration, in der ebenfalls Soundex als Algorithmus eingesetzt wurde, da die Hinzunahme der Postleitzahl keine Auswirkungen auf die Anzahl der Verknüpfungsfehler hatte.
Somit konnte das Kriterium „Multiple-Value-Felder", aus denen sich die Homonymfehler ergaben, sowie das Kriterium „Tippfehler", aus dem sich der Synonymfehler ergab, nicht berücksichtigt werden.
Auch das zusätzliche Kriterium der Adressänderung konnte nicht korrekt berücksichtigt werden.

Die sechste Konfiguration verwendete erneut den deterministischen Abgleich. In diesem Fall führte die Hinzunahmen der fünften Matching-Variable zu einer Verringerung der Verknüpfungsfehler.
Im Vergleich zu Konfiguration 3, in der die Postleitzahl nicht als Matching-Variable verwendet wurde, traten in diesem Test keine Homonymfehler auf.
Die Anzahl der Synonymfehler blieb unverändert und resultierte erneut aus Tippfehlern.
Somit konnte das Kriterium „Multiple-Value-Felder" ebenso wie das Kriterium der Adressänderung berücksichtigt werden. Eine Berücksichtigung der Tippfehler konnte unter Verwendung dieser Konfiguration jedoch nicht vollständig erfolgen.

Die Integration der Postleitzahl als zusätzliche Matching-Variable in diesem Durchlauf führte lediglich im Kontext des deterministischen Abgleichs zu einer Verbesserung der Verknüpfungsqualität.
Die Anzahl der Verknüpfungsfehler in den Konfigurationen 4 und 5 konnte durch die Einbeziehung der Postleitzahl nicht reduziert werden, da hier die korrekte Identifizierung weitere Fehler zur Folge gehabt hätte, was wiederum die gesamte Verknüpfungsqualität beeinträchtigt hätte.
Die Konfiguration 4 konnte unter Verwendung der Levenshtein-Distanz erneut die beste Verknüpfungsqualität erzielen.

4.2.3.3. Test-Durchlauf 3

Im Rahmen der Test-Durchläufe 3 und 4 wurde der Goldstandard-Datensatz des Krebsregisters durch die Integration weiterer Testdatensätze von 7510 auf 7527 Datensätze erhöht.

FRIL Durchlauf 3	**Algorithmus**	**TP**	**TN**	**FP**	**FN**	**Recall**	**Precision**	**F-Score**
Konfiguration 7	Levenshtein-Distanz	1258	6264	2	3	0,9976	0,9984	0,9980
Konfiguration 8	Soundex	1259	6263	3	2	0,9984	0,9976	0,9980
Konfiguration 9	Deterministisch	1251	6264	2	10	0,9921	0,9984	0,9952

Tabelle 4.9.: Testergebnisse von FRIL für Test-Durchlauf 3 (entnommen aus: eigene Aufnahmen)

Konfiguration 7 setzte erneut die Levenshtein-Distanz ein und wies im Vergleich zur Konfiguration 1, die auf den Originaldatensatz angewendet wurde, drei neue Synonymfehler auf. Die drei Synonymfehler ergaben sich aus einem Tippfehler, einer Nachnamensänderung und einem Zahlendreher.
Die zwei entstandenen Homonymfehler resultierten, analog zu denen der ersten Konfiguration, aus Fällen, die sich lediglich in der Postleitzahl unterschieden.
Trotz erfolgreicher Identifikation der Kriterien Zwillingspaar, Multiple-Value-Felder, unterschiedliche Verwendung von Diakritika und die Geschwister, die am gleichen Tag Geburtstag haben, konnten Nachnamensänderungen und Tippfehler nicht korrekt berücksichtigt werden.

Konfiguration 8 nutzte Soundex als Algorithmus und zeigte trotz der Hinzunahme weiterer Testfälle eine geringere Anzahl an Fehlern im Vergleich zu Konfiguration 2, bei der Soundex auf den Originaldatensatz angewendet wurde.
Die drei Homonymfehler bestanden aus Fällen, in denen lediglich die Postleitzahl unterschiedlich war sowie aus einem Fall, in dem zwei Datensätze einen ähnlichen Nachnamen hatten, aber in allen weiteren Matching-Variablen Unterschiede aufwiesen. Die zwei Synonymfehler resultierten aus einem Tippfehler und einer Änderung des Nachnamens.

Die neunte Konfiguration verwendet einen deterministischen Abgleich und zeigte aufgrund der Hinzunahme weiterer Testfälle eine erhöhte Anzahl an Verknüpfungsfehlern im Vergleich zur Konfiguration 3 im ersten Test.
Die zwei entstandenen Synonymfehler resultierten erneut aus Fällen, die sich lediglich in ihrer Postleitzahl unterschieden.
Mit zehn Homonymfehlern wies diese Konfiguration die bisher höchste Anzahl dieser Fehler auf. Diese Homonymfehler sind aus Tippfehlern, Nachnamensänderungen, Zahlendrehern und der unterschiedlichen Verwendung von Diakritika entstanden. Alle diese Kriterien konnten mit der neunten Konfiguration nicht erfüllt werden, während die Kriterien „Zwillingspaar", „Multiple-Value-Felder" und „Geschwister die am gleichen Tag Geburtstag haben" korrekt identifiziert wurden.

Die Hinzunahme weiterer Testfälle führte, wie erwartet in einigen Tester-

gebnissen zu einer Zunahme der Verknüpfungsfehler.
In diesem Testdurchlauf weisen Soundex und die Levenshtein-Distanz eine gleich hohe Verknüpfungsqualität auf. Wenn die Anzahl der Synonymfehler priorisiert wird, hat die Konfiguration unter Verwendung von Soundex jedoch die höchste Verknüpfungsqualität.

4.2.3.4. Test-Durchlauf 4

Im Rahmen des Tests 4 wurde die Postleitzahl erneut als zusätzliche Matching-Variable hinzugezogen und die entwickelten Konfigurationen auf den erweiterten Goldstandard-Datensatz angewendet.

FRIL Durchlauf 4	**Algorithmus**	**TP**	**TN**	**FP**	**FN**	**Recall**	**Precision**	**F-Score**
Konfiguration 10	Levenshtein-Distanz	1258	6264	2	3	0,9976	0,9984	0,9980
Konfiguration 11	Soundex	1259	6263	3	2	0,9984	0,9976	0,9980
Konfiguration 12	Deterministisch	1250	6266	0	11	0,9913	1	0,9956

Tabelle 4.10.: Testergebnisse von FRIL für Test-Durchlauf 4 (entnommen aus: eigene Aufnahmen)

Die Ergebnisse der Konfiguration 10, die die Levenshtein-Distanz verwendete, zeigten keinen signifikanten Unterschied in der Anzahl der Verknüpfungsfehler im Vergleich zur Konfiguration 7 auf, in der die Postleitzahl nicht verwendet wurde.
Daher konnten erneut Zwillingspaare, Multiple-Value-Felder, die unterschiedliche Verwendung von Diakritika und Geschwister, die am gleichen Tag Geburtstag haben, korrekt identifiziert werden, während Nachnamensänderungen, Adressänderungen und Tippfehler nicht vollständig berücksichtigt werden konnten.

Die Konfiguration 11 wies ebenfalls keine wesentlichen Unterschiede in der Anzahl der Verknüpfungsfehler im Vergleich zur Konfiguration 8 auf.
Die beiden Homonymfehler, bei denen sich die Datensätze lediglich in der Postleitzahl unterscheiden, konnten auch mit dieser Konfiguration nicht korrekt identifiziert werden.
Der dritte Homonymfehler ergibt sich in diesem Testdurchlauf im Gegensatz zur achten Konfiguration nicht aufgrund eines ähnlichen Nachnamens, sondern infolge eines Multiple-Value-Felds.
Die Ursachen für die Synonymfehler blieben unverändert.
Folglich konnten Zwillingspaare, die unterschiedliche Verwendung von Diakritika und Geschwister, die am gleichen Tag Geburtstag haben, korrekt identifiziert werden, während Nachnamensänderungen, Multiple-Value-Felder, Adressänderungen und Tippfehler nicht vollständig berücksichtigt werden konnten.

In der zwölften Konfiguration, die einen deterministischen Abgleich implementierte, sind keine Homonymfehler entstanden. Jedoch ist die Anzahl der Synonymfehler durch die Hinzunahme der Postleitzahl von zehn in Konfiguration 9 auf elf in Konfiguration 12 angestiegen. Die erfolgreiche Vermeidung von Homonymfehlern ermöglichte eine korrekte Zuordnung der Fälle, in denen ausschließlich eine Differenz in der Postleitzahl vorhanden war.
Dennoch zeigt sich, dass die Anzahl der Synonymfehler in dieser Konfiguration mit elf Fehlern im Vergleich zu den Konfigurationen 10 und 11 signifikant höher ausfiel. Das unterstreicht erneut die geringe Berücksichtigung der Verknüpfungskriterien bei einem deterministische Abgleich.
Insbesondere konnten in der zwölften Konfiguration keine Tippfehler, Nachnamensänderungen, Zahlendreher, Adressänderungen und unterschiedliche Verwendungen von Diakritika adressiert werden. Hingegen wurden Zwillinge, Multiple-Value-Felder und Geschwister mit gleichem Geburtsdatum korrekt identifiziert.
Die Hinzunahme der Postleitzahl führte in diesem Test zu keiner signifikanten Veränderung der Verknüpfungsqualität. In diesem Testdurchlauf weisen Soundex und die Levenshtein-Distanz erneut eine gleich hohe Verknüpfungsqualität auf. Wenn die Anzahl der Synonymfehler priorisiert wird, zeigt die Konfiguration unter Verwendung von Soundex die höchste Verknüpfungsqualität.

4.2.4. Gegenüberstellung Verknüpfungsqualität

Ergebnisse	Recall E-PIX	Recall FRIL	Precision E-PIX	Precision FRIL	F-Score E-PIX	F-Score FRIL	Datensatz	Algorithmus	Matching-Variablen
Konfiguration 1	1	1	0,9984	0,9984	0,9992	0,9992	Original	Levenshtein-Distanz	Pflichtvariablen
Konfiguration 2	0,9984	0,9992	0,9984	0,9952	0,9984	0,9972	Original	Kölner Phonetik	Pflichtvariablen
Konfiguration 3	0,9976	0,9976	0,9984	0,9984	0,9980	0,9980	Original	Deterministisch	Pflichtvariablen
Konfiguration 4	0,9992	1	1	0,9984	0,9996	0,9992	Original	Levenshtein-Distanz	Pflichtvariablen + PLZ
Konfiguration 5	0,9984	0,9992	0,9984	0,9952	0,9984	0,9972	Original	Kölner Phonetik	Pflichtvariablen + PLZ
Konfiguration 6	0,9968	0,9968	1	1	0,9984	0,9984	Original	Deterministisch	Pflichtvariablen + PLZ
Konfiguration 7	0,9976	0,9976	0,9968	0,9984	0,9972	0,9980	Erweitert	Levenshtein-Distanz	Pflichtvariablen
Konfiguration 8	0,9960	0,9984	0,9984	0,9976	0,9972	0,9980	Erweitert	Kölner Phonetik	Pflichtvariablen
Konfiguration 9	0,9921	0,9921	0,9984	0,9984	0,9952	0,9952	Erweitert	Deterministisch	Pflichtvariablen
Konfiguration 10	0,9992	0,9976	0,9976	0,9984	0,9984	0,9980	Erweitert	Levenshtein-Distanz	Pflichtvariablen + PLZ
Konfiguration 11	0,9980	0,9984	0,9984	0,9976	0,9968	0,9980	Erweitert	Kölner Phonetik	Pflichtvariablen + PLZ
Konfiguration 12	0,9913	0,9913	1	1	0,9956	0,9956	Erweitert	Deterministisch	Pflichtvariablen + PLZ

Tabelle 4.11.: Übersicht der Verknüpfungsqualität der einzelnen Konfigurationen von E-PIX und FRIL [PLZ = Postleitzahl](entnommen aus: eigene Aufnahmen)

Im Folgenden werden die Verknüpfungsergebnisse der einzelnen Testdurchläufe des E-PIX und FRIL gegenübergestellt.
Die Ergebnisse in Tabelle 4.11 weisen auf eine durchgängig hohe Verknüpfungsqualität für beide Lösungen hin. Sämtliche Werte für Recall, Precision und F-Maß liegen in allen Testdurchläufen zwischen 0,99 und 1.
Die Verknüpfungsergebnisse unterliegen neben der Wahl des Algorithmus weiteren Einflussfaktoren. Hierzu zählen insbesondere die Festlegung der Gesamtschwellwerte und die Konfiguration der Gewichte sowie Schwellwerte für die Matching-Variablen.

Die hohen Werte des Recalls deuten darauf hin, dass beide Lösungen effektiv in der Lage sind, zusammengehörige Datensätze zu identifizieren und einen Großteil der tatsächlich zusammengehörigen Datensätze erfolgreich erkennen können.
Die ebenfalls hohen Werte für die Precision zeigen, dass die von den Record Linkage Lösungen identifizierten zusammengehörigen Datensätzen in über 99% aller Fälle tatsächlich korrekt sind und zusammengehören. Dies deutet darauf hin, dass E-PIX und FRIL die Datensätze präzise verknüpfen und lediglich eine minimale Anzahl an falsch positiven Ergebnissen (Homonymfehler) generieren.
Da sich sowohl der Recall als auch die Precision in einem Bereich zwischen 0,99 und 1 bewegen, liegt ein relativ ausgewogenes Verhältnis zwischen der Identifikation zusammengehöriger Datensätze und der Vermeidung falsch positiver Ergebnisse vor. Dies wird mit der Betrachtung des F-Maß deutlich. In einigen Fällen zeigt sich unter Anwendung des E-PIX eine höhere Ausgewogenheit, während in anderen Fällen unter Verwendung von FRIL eine höhere Ausgewogenheit zu beobachten ist, jedoch gleichen sich diese Unterschiede aus.
Bei der Entwicklung der Konfigurationen wurde die Minimierung von Synonymfehlern gegenüber Homonymfehlern priorisiert. Daher ist die Metrik des Recalls besonders bedeutsam, denn ein hoher Recall weist darauf hin, dass die Record Linkage Lösung dazu tendiert, zumindest annähernd alle zusammengehörigen Datenpaare zu identifizieren. Ein niedriger Recall würde darauf hindeuten, dass einige zusammengehörige Datenpaare übersehen wurden. Ein Vergleich der Ergebnisse des Recalls zeigt, dass FRIL in der Mehrzahl der Tests einen minimal besseren Recall erzielte als der E-PIX. Dagegen schneidet der E-PIX bei den Ergebnissen für die Precision minimal besser als FRIL ab. In Konfiguration 10 zeigt sich der größte Unterschied im Recall, wobei jedoch der E-PIX mit einen Wert von 0,9992 einen höheren Recall aufweist im Vergleich zu FRIL, dessen Wert bei 0,9976 liegt.

Wird die Vermeidung von Synonymfehlern priorisiert, erzielen beide Record Linkage Lösungen insgesamt die höchste Verknüpfungsqualität bei der

Verwendung der Levenshtein-Distanz.
Im Rahmen von FRIL weisen stets die Konfigurationen die gerinste Verknüpfungsqualität auf, die entweder auf Basis des Soundex oder auf Grundlage des deterministischen Algorithmus beruhen, während mit dem E-PIX der deterministische Vergleich stets die geringste Verknüpfungsqualität aufweist. Hierbei ist zu beachten, dass mit FRIL die niedrigste Verknüpfungsqualität des deterministsichen Abgleichs geringer ist als die niedrigste Verknüpfungsqualität des Soundex.

Die höchste Verknüpfungsqualität zeigt sich in Konfiguration 4 des E-PIX, die auf den Originaldatensatz mit der zusätzlichen Matching-Variable „Postleitzahl" angewendet wurde.
Hinsichtlich der Testergebnisse des erweiterten Datensatzes (Konfigurationen 7-12) konnte mit Konfiguration 10, unter Einbeziehung der Postleitzahl, die höchste Verknüpfungsqualität sowohl mit FRIL als auch mit dem E-PIX erreicht werden. Hierbei zeigten beide Lösungen insgesamt fast gleichwertige Verknüpfungsqualitäten.
Mit einem Recall von 0,9992 und einem F-Score von 0,9984 zeigt der E-PIX eine minimal höhere Verknüpfungsqualität als FRIL mit einem Recall von 0,9976 und einem F-Score von 0,9980.

Eine Überprüfung mit dem Krebsregister Mecklenburg-Vorpommern, das den E-PIX für das Recod-Linkage nutzt, hat ergeben, dass einige Synonymfehler, die sich auf Personen mit mehreren Identitäten beziehen, zuvor manuell zusammengefügt werden mussten. Mit der Anwendung der Konfiguration 4 auf den Originaldatensatz aus dem Krebsregister und deren Matching-Variablen, konnten diese Datensätze automatisch korrekt zusammengefügt werden. Daraus folgt, dass mit dieser Konfiguration weniger Datensätze manuell hätten zusammengeführt werden müssen.
Für Register, die die Postleitzahl als fünfte Matching-Variable verwenden, erweist sich Konfiguration 10 des E-PIX als empfehlenswert. Diese Konfiguration wurde auf den erweiterten Datensatz angewendet, wodurch zusätzliche potenzielle Fehlerfälle berücksichtigt wurden. Somit konnte eine Konfiguration, die als Grundlage für Register dienen kann, entwickelt und veröffentlicht werden, von der weitere Register profitieren können[THS-Community, 2024].
Für Register, bei denen die Postleitzahl nicht als Variable verfügbar ist, kann die Nutzung von Konfiguration 7 als Grundlage in Betracht gezogen werden. Gleiches gilt für die Konfigurationen 7 und 10 von FRIL.

Zusammenfassend lässt sich festhalten, dass mit beiden Record Linkage Lösungen eine sehr hohe Verknüpfungsqualität erzielt werden konnte. Außerdem konnten Unterschiede in der Verknüpfungsqualität der beiden Lösungen festgestellt werden, die sich jedoch insgesamt ausgleichen.
Im Allgemeinen kann die Levenshtein-Distanz als Algorithmus für den Vergleich von Matching-Variablen empfohlen werden, da mit beiden Reocrd-Linkage Lösungen die besten Ergebnisse unter Verwendung dieses Algorithmus erzielt wurden.

Um einen detaillierteren Einblick in die entstandenen Verknüpfungsfehler der Record Linkage Lösungen zu bekommen, wird im folgenden Balkendiagramm der prozentualen Anteile der Verknüpfungsfehler jeder Konfiguration zusammengefasst dargestellt.

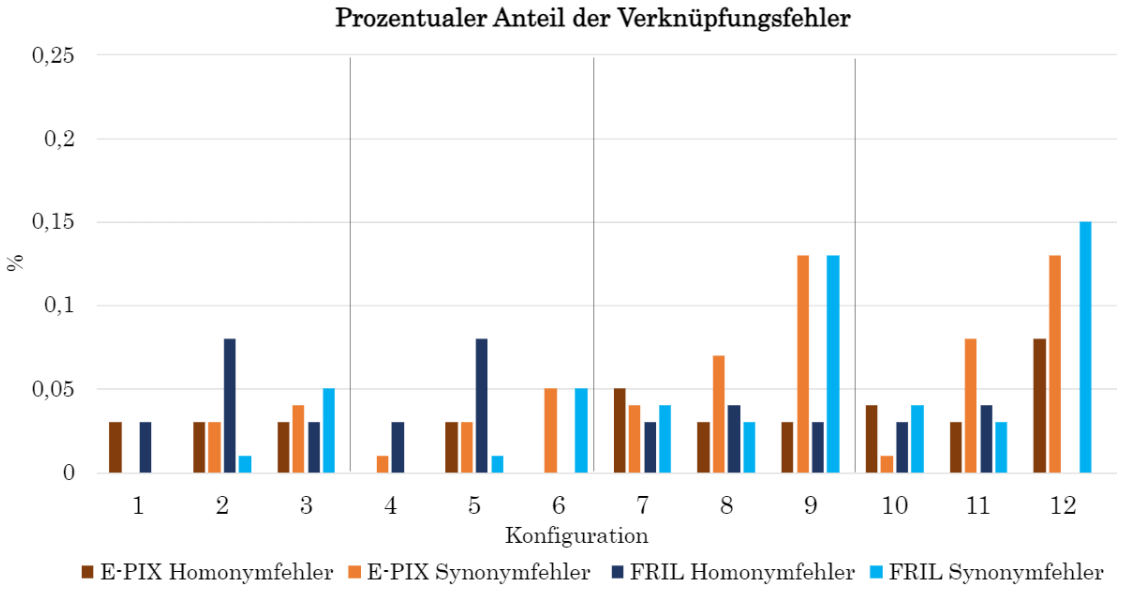

Abbildung 4.3.: Prozentualer Fehleranteil für jede Konfiguration des E-PIX und FRIL (entnommen aus: eigene Aufnahmen)

Die grafische Darstellung in Form eines Balkendiagramms präsentiert den prozentualen Fehleranteil für jede getestete Konfiguration der beiden Record Linkage Lösungen. Die vertikale Achse repräsentiert den Prozentsatz, während auf der horizontalen Achse die einzelnen Konfigurationen aufgeführt sind. Die orangefarbenen Balken stellen den prozentualen Anteil der Verknüpfungsfehler des E-PIX und die blau gefärbten Balken die Verknüpfungsfehler von FRIL dar.
Die Grafik verdeutlicht, dass der prozentuale Anteil der Verknüpfungsfehler in allen Konfigurationen mit dem E-PIX und FRIL weit unter einem Prozent liegt,

was erneut auf eine insgesamt sehr gute Verknüpfungsqualität hinweist.

Insbesondere zeigt sich für den E-PIX, dass die Verwendung der Levenshtein-Distanz in den Konfigurationen eins, vier und zehn dazu beiträgt, Synonymfehler effektiv zu minimieren. In den Konfigurationen des E-PIX, die nicht die Levenshtein-Distanz als Algorithmus nutzen, machen Synonymfehler einen größeren Teil der Verknüpfungsfehler aus.
Die Erweiterung des Goldstandard-Datensatzes ab Konfiguration 7 führt zu einer erkennbaren Zunahme der Verknüpfungsfehler in den Konfigurationen des E-PIX, was auf die Integration zusätzlicher schwer zu identifizierender Datensätze zurückzuführen ist.
Des Weiteren zeigt das Diagramm, dass die Verwendung der Postleitzahl als fünfte Matching-Variable zu einer Reduzierung der Verknüpfungsfehler des E-PIX führt. Dies wird insbesondere in den Konfigurationen 1, 4, 7 und 10 deutlich. Konfigurationen 1 und 7, die keine Postleitzahl als zusätzliche Variable berücksichtigen, weisen höhere Fehleranteile auf, als die Konfigurationen 4 und 10, die die Postleitzahl einbeziehen.

Mit Blick auf den prozentualen Fehleranteil der Konfigurationen im Rahmen von FRIL wird deutlich, dass die Verwendung der Levenshtein-Distanz in den Konfigurationen 1, 4, 7 und 10 wie auch beim E-PIX zu den geringsten prozentualen Fehleranteilen führt und lediglich nur eine geringe Anzahl an Synonymfehler auftreten.
Weiter kann beobachtet werden, dass die Konfigurationen 1 und 4 sämtliche Synonymfehler identifizieren konnten, gefolgt von den Konfigurationen 2 und 5, die Soundex als Algorithmus nutzen und ebenfalls nur einen geringen Anteil an Synonymfehlern aufweisen.
In den ersten beiden Testdurchläufen des Originaldatensatz (Konfigurationen 1 bis 6) überwiegt in nahezu allen Konfigurationen von FRIL die Anzahl der Homonymfehler. Mit der Integration weiterer Testfälle steigt erwartungsgemäß, wie es auch beim E-PIX zu beobachten ist, die Anzahl der Synonymfehler, wobei diese häufig im Vergleich zu den Homonymfehlern einen größeren Anteil der Verknüpfungsfehler ausmachen.
Deutlich wird ebenfalls, dass die Hinzunahme der Postleitzahl bei der Verwendung von FRIL lediglich eine Auswirkung auf die Verringerung der Synonymfehler bei einem deterministischen Abgleich hatte und nicht wie unter Verwendung des E-PIX in mehreren Konfigurationen zu einer Reduzierung der Fehler führte.
Dies wir deutlich anhand des Vergleichs zwischen den Konfigurationen 3 und 6, die beide auf den Originaldatensatz angewendet wurden, aber Konfiguration 6 die Postleitzahl als weitere Matching-Variable verwendete. Auch die Betrachtung

der Konfigurationen 9 und 12 verdeutlicht, dass unter Verwendung der Postleitzahl in Konfiguration 12 weniger Synonymfehler vorliegen als in Konfiguration 9.

In einem nächsten Schritt werden die Record Linkage Lösungen hinsichtlich der Berücksichtigung der Verknüpfungskriterien verglichen.

Kriterium/Algorithmus	Levenshtein-Distanz		Soundex		Deterministisch		*Legende:*
Tippfehler	+	-	-	-	-	-	E-PIX
Zahlendreher	+	+	+	+	-	-	
Zwillingspaare	+	+	+	+	+	+	FRIL
Adressänderung	-	-	-	-	-	-	
Multiple-Value-Felder	+	+	+	-	+	+	
Nachnamensänderung	-	-	-	-	-	-	
Verwendung von Diakritika	+	+	+	+	-	-	
Geschwister mit gleichem Geburtstag	+	+	+	+	+	+	

Tabelle 4.12.: Berücksichtigung (+) und zum Teil Berücksichtigung (-) der Verknüpfungskriterien mit E-PIX und FRIL (entnommen aus: eigene Aufnahmen)

Die Zusammenfassung der Resultate in Tabelle 4.12 zeigt, dass mit dem E-PIX unter Verwendung der Levenshtein-Distanz sechs von acht Verknüpfungskriterien erfolgreich berücksichtigt werden konnten, jedoch zwei Kriterien nur zum Teil erfüllt wurden, gleichzeitig konnte mit diesem Algorithmus die höchste Verknüpfungsqualität erzielt werden. Die Kölner Phonetik erreichte unter Verwendung des E-PIX die zweitbesten Ergebnisse, indem fünf von acht Verknüpfungskriterien vollständig berücksichtigt wurden. Mit einem deterministsiche Abgleich, der die niedrigste Verknüpfungsqualität erzielte, konnten lediglich drei Kriterien erfüllt werden. Es ist zu beachten, dass die Verantwortung für die Einhaltung der Kriterien nicht allein beim Algorithmus selbst liegt. Entscheidend sind auch die gewählten Schwellenwerte und Gewichtungen, die auf Basis des Fehlerabgleichs mit dem Referenzdatensatz festgelegt und angepasst wurden.
Die Reaktion des E-PIX auf Multiple-Value-Felder spielte in allen Konfigurationen eine entscheidende Rolle und trug maßgeblich zur Erhöhung der Verknüpfungsqualität bei, da dieses Kriterium bei Verwendung aller Algorithmen erfüllt wurde.
Die Kölner Phonetik berücksichtigt unter Verwendung des E-PIX weniger Kriterien als die Levenshtein-Distanz, was der Funktionsweise dieses Algorithmus geschuldet ist. Datensätze mit teilweise unterschiedlichen Werten trotz gleicher Identität erhalten wie in Kapitel 2.2.3 erläutert, unterschiedliche Codes, was insbesondere bei Adress- und Nachnamensänderungen sowie Tippfehlern relevant ist.
Die Evaluation des Verknüpfungsprozess mittels FRIL zeigt, dass die Verwendung der Levenshtein-Distanz, wie auch bei der Verwendung des E-PIX, die meisten Ver-

knüpfungskriterien berücksichtigt und gleichzeitig die höchste Verknüpfungsqualität erzielte. Die Kriterien „Tippfehler", „Adressänderung" und „Nachnamensänderung" wurden mit der Levenshtein-Distanz in FRIL nur teilweise berücksichtigt, was zu Verknüpfungsfehlern geführt hat. Es ist anzumerken, dass eine Berücksichtigung des Kriteriums „Adressänderung" aufgrund der Charakteristik des vorliegenden Datensatzes zu einer Erhöhung der Synonymfehler geführt hätte und dieses Kriterium somit aufgrund der Priorisierung der Synonymfehler nicht berücksichtigt werden konnte. Jedoch ist zu betonen, dass fünf von acht Kriterien unter Verwendung dieses Algorithmus erfolgreich berücksichtigt wurden.
FRIL konnte mit Soundex weniger Kriterien erfüllen als mit der Levenshtein-Distanz. Mit Soundex konnte lediglich die Hälfte der Kriterien berücksichtigt werden, wobei Tippfehler, Adressänderungen, Nachnamensänderungen und Multiple-Value-Felder zu Verknüpfunsgfehlern führten. Mit dem deterministischen Abgleich konnte die Mehrheit der Verknüpfungskriterien in FRIL nicht zuverlässig berücksichtigt werden. Dies erklärt die teilweise geringste Verknüpfungsqualität, die in den Konfigurationen unter Verwendung des determinstsichen Abgleichs auftrat. Es ist wichtig zu beachten, dass bei einem deterministischen Vergleich nicht nur die Funktionsweise des Algorithmus selbst, sondern auch die gewählten Gewichtungen der Matching-Variablen, basierend auf dem Abgleich von Fehlern mit dem Referenzdatensatz, einen Einfluss auf das Endergebnis haben. Lediglich drei der acht Verknüpfungskriterien konnten mit dieser Methode zuverlässig erfüllt werden.
Aus der vorliegenden Tabelle wird deutlich, dass sowohl der E-PIX als auch FRIL unter Verwendung der Levenshtein-Distanz die meisten Kriterien berücksichtigen konnten. Die Verwendung dieses Algorithmus führte in beiden Record Linkage Lösungen ebenfalls zu der höchsten Verknüpfungsqualität, was darauf hinweist, dass die präzise Berücksichtigung der Verknüpfungskriterien mit einer höheren Verknüpfungsqualität einhergeht. Dabei ist anzumerken, dass der E-PIX mit seinen Konfigurationen ein zusätzliches Kriterium im Vergleich zu FRIL berücksichtigen konnte.
Die Anwendung von Soundex ermöglichte beiden Lösungen die Berücksichtigung der zweitmeisten Kriterien, während unter Verwendung des deterministischen Abgleichs sowohl E-PIX als auf FRIL die geringste Anzahl an Kriterien erfüllen konnten. Mit Ausnahme der Kriterien „Tippfehler" und „Multiple-Value-Felder", die von FRIL nicht vollständig eingehalten werden konnten, zeigt sich unter Verwendung der Levenshtein-Distanz und Soundex eine Übereinstimmung in den nicht berücksichtigten Kriterien der beiden Record Linkage Lösungen.

Die teilweise falsche Identifikation von Multiple-Value-Feldern durch FRIL könnte auf das Fehlen einer allgemeinen Konfigurationsmöglichkeit, mit der auf

diese Felder angemessen reagiert werden kann, zurückzuführen sein. Während die Verwendung der Levenshtein-Distanz in FRIL eine korrekte Identifikation dieser Felder ermöglicht, scheint die Implementierung von Soundex diese Möglichkeit einzuschränken. Im Gegensatz dazu gewährleistet E-PIX durch seine Konfigurationsmöglichkeit, auf Multiple-Value-Felder zu reagieren, unabhängig vom gewählten Algotihmus stets die Berücksichtigung dieses Kriteriums.

5. Fazit

5.1. Zusammenfassung der Ergebnisse

Eine Forschungsfrage der vorliegenden Arbeit zielt auf die Einflussfaktoren der Entstehung von Verknüpfungsfehlern ab. Die Analysen zeigen, dass die Qualität der Verknüpfung von unterschiedlichen Faktoren beeinflusst wird, die in die Kategorien Datenqualität und technische Realisierung unterteilt werden können.
Im Bereich der Datenqualität spielen Faktoren wie etwa die Verwendung von Quais-Identifikatoren, unterschiedliche Methoden der Dateneingabe, die Nutzung verschiedener Datenformate, Tippfehler, Zahlendreher, motorische Fehler, verschiedene Variationen von Wörtern sowie Herausforderungen bei Multiple-Value-Feldern und dynamischen Variablen eine Rolle.
Die technische Umsetzung des Record Linkage prozess beeinflusst ebenfalls die Verknüpfungsqualität. Eine breite Palette von Konfigurierungsmöglichkeiten innerhalb der Record Linkage Lösungen ist von besonderer Bedeutung, um den Verknüpfungsprozess detailliert anzupassen und spezifisch auf einen Datensatz zuzuschneiden, um so die Anzahl der Fehler minimieren zu können. Ein Beispiel hierfür ist die Konfigurierbarkeit des E-PIX bei der Reaktion auf Multiple-Value-Felder. Das Fehlen eines einheitlich Goldstandard-Datensatzes behindert derzeit die Evaluation von Record Linkage Lösungen und beeinträchtigt daher indirekt ebenfalls die Qualität der Datensatzverknüpfung.

Mit Blick auf das Forschungsinteresse, die Konsequenzen von Verknüpfungsfehlern zu analysieren, zeigen die Ergebnisse dieser Arbeit, dass eine unzureichende Verknüpfungsqualität schwerwiegende Fehler wie Selektionsverzerrungen, Fehlklassifizierungen, Informationsverzerrungen und Messfehler verursachen kann. Diese Folgen können erhebliche Auswirkungen sowohl auf die Forschung als auch auf die menschliche Gesundheit haben, insbesondere wenn sie dazu führen, das Risiko bestimmter Krankheiten, wie beispielsweise Krebs, zu unterschätzen.

Zur Beantwortung der Forschungsfrage, inwieweit bestehende Record Linkage Lösungen die Verknüpfungsqualität durch die Berücksichtigung wesentlicher Kriterien steigern können, wurden im Rahmen einer Marktanalyse elf Record Linkage Lösungen kurz analysiert und schließlich die Lösungen E-PIX und FRIL für eine empirische Untersuchung ausgewählt.
Die empirische Untersuchung zeigt, dass beide Record Linkage Lösungen bei einer Anwendung der Levenshtein-Distanz eine Vielzahl von Verknüpfungskriterien berücksichtigen können. Sie weisen in dieser Konfiguration im Vergleich zur

Nutzung anderer Algorithmen auch die höchste Verknüpfungsqualität auf.
Zu beachten ist jedoch, dass neben der Wahl des Algorithmus weitere Faktoren einen Einfluss auf die Verknüpfungsqualität haben. Hierzu zählen insbesondere die Festlegung der Gesamtschwellwerte und die Konfiguration der Gewichte sowie Schwellwerte für die Matching-Variablen. Sowohl E-PIX als auch FRIL zeigten Schwächen im Umgang mit dynamischen Variablen wie Adressänderungen und Nachnamensänderungen. Beide Lösungen reagierten teilweise auch nicht korrekt auf Tippfehler.FRIL stieß insbesondere auf Schwierigkeiten bei der korrekten Identifizierung von Multiple-Value-Feldern. Dies lässt sich darauf zurückführen, dass hier im Vergleich zum E-PIX eine zielführende Konfigurationsmöglichkeit fehlt.
Es ist hervorzuheben, dass sämtliche Tests für beide Lösungen eine sehr hohe Verknüpfungsqualität mit einem Recall und einer Precision zwischen 0,99 und 1 aufwiesen, was auf eine insgesamt sehr hohe Verknüpfungsqualität schließen lässt. Darüber hinaus wurden Unterschiede in der Höhe der Verknüpfungsqualität zwischen den beiden Lösungen festgestellt, die sich jedoch insgesamt ausgleichen.
Ein zusätzliches Ergebnis der empirischen Analyse ist, dass sich Konfiguration 10 sehr gut als Grundlage für Register eignet. Diese Konfiguration wurde öffentlich zugänglich gemacht, sodass andere Register von der Konfiguration profitieren können[THS-Community, 2024].

5.2. Diskussion

Die vorliegenden Ergebnisse zeigen eine Reihe von Einflussfaktoren der Datenqualität und der technischen Umsetzung auf die Verknüpfungsqualität und somit auf die Entstehung von Verknüpfungsfehlern auf. Der festgestellte Zusammenhang zwischen Datenqualität, technischer Umsetzung und Verknüpfungsqualität untermauert frühere wissenschaftliche Veröffentlichungen, die bereits auf diesen Zusammenhang hingewiesen haben.
Die herausgearbeiteten Konsequenzen von Verknüpfungsfehlern zeigen deutlich, dass eine hohe Verknüpfungsqualität von essenzieller Bedeutung ist, um schwerwiegende und potenziell gefährliche Folgen zu vermeiden. Diese Erkenntnis steht im Einklang mit anderen Publikationen, die bereits auf weitreichenden Konsequenzen von Verknüpfungsfehlern für Forschungsergebnisse und für die menschliche Gesundheit hingewiesen haben.
Fehlklassifizierungen, Selektionsfehler und Messfehler, die durch eine Verknüpfung entstehen können, müssen in Zukunft gezielter vermieden werden, um korrekte Forschungsergebnisse zu gewährleisten und potenzielle Risiken für die Gesundheit zu minimieren. In diesem Kontext wurden in dieser Arbeit die Record Linkage Lösungen E-PIX und FRIL evaluiert und es wurde eine Empfehlung für eine geeig-

nete Konfiguration in Registern entwickelt, mit dem Ziel, die Verknüpfungsqualität in Registern zukünftig zu verbessern.
Es ist jedoch zu betonen, dass neben E-PIX und FRIL, eine Vielzahl weiterer Record Linkage Lösungen existieren, deren Verknüpfungsqualität ebenfalls evaluiert werden sollte, um diese gegebenenfalls weiterzuentwickeln. Zudem ist es erforderlich, über den Anwendungsbereich von Registern hinaus weitere geeignete Konfigurationen zu entwickeln und diese zu veröffentlichen. Dies ist wichtig, um sicherzustellen, dass neu geschaffene Datenressourcen effektiv genutzt werden.

Die Erstellung und Nutzung von Goldstandard-Datensätzen stellen eine große Herausforderung in der Record Linkage Forschung dar. In dieser Arbeit konnte mittels des Zugriffs auf IDAT ein Goldstandard-Datensatz verwendet werden, der Echtdaten enthält und somit als Referenz für den Record Linkage Prozess dient. Es ist jedoch zu beachten, dass trotz des Abgleichs mit einem Melderegister falsche Angaben über die tatsächlichen Verknüpfungen nicht gänzlich ausgeschlossen werden können, auch wenn ein solcher Abgleich eine äußerst sichere Variante für die Erstellung eines Goldstandard-Datensatzes ist.
Um neben den enthaltenden Fehlerfällen, die in dem Goldstandard-Datensatz vorliegen auch weitere Kriterien zu prüfen, wurden in weiteren Tests fiktive Datensätze hinzugefügt. In diesem Szenario ist zu berücksichtigen, dass die hinzugefügten Datensätze unter Umständen nicht vollständig der Charakteristik von Echtdaten entsprechen.

Die aktuelle Herausforderung bei Studien, die die Präzision von Record Linkage Lösungen messen oder verschiedene Algorithmen testen wollen, liegt in der Schwierigkeit, einen Referenzdatensatz zu finden.
Ein standardisierter Testdatensatz für dieses Forschungsgebiet könnte einen aussagekräftigeren Vergleich zwischen verschiedenen Record Linkage Lösungen ermöglichen und dazu beitragen, Optimierungsbereiche zu identifizieren und Verknüpfungsfehler zu verhindern. Dieser Schritt wäre entscheidend, um die Qualität von Record Linkage Lösungen zu verbessern und eine einheitliche Grundlage zur Evaluierung zu schaffen.
Die größten Herausforderungen, die die Verknüpfungsqualität im Bereich des Record Linkage beeinflussen, liegen insbesondere in der Handhabung von Quais-Identifikatoren. Die Verwendung dieser Identifikatoren führt häufig zu Fehlern in der Datenqualität, was die Genauigkeit des Verknüpfungsprozess beeinträchtigt. Der Entwicklung eines eindeutigen Identifikators stehen allerdings erhebliche Hindernisse entgegen, darunter Datenschutzbedenken und ethische Fragestellungen.

Angesichts dieser Herausforderungen erscheint es notwendig, den Austausch von Problemstellungen und hierfür geeignete Lösungen zu intensivieren. Durch verstärkte Kooperation könnte eine gemeinsame Nutzung von Konfigurationen sowie der Austausch über Herausforderungen zu einer Minimierung von Verknüpfungsfehlern führen.
Ein Schritt in diese Richtung wurde mit dem publizierten White Paper zur Verbesserung des Record Linkage für die Gesundheitsforschung in Deutschland gemacht, in dem Anwendungsfälle des Record Linkage in der Gesundheitsforschung mit ihren Herausforderungen vorgestellt sowie Lösungswege aufgezeigt wurden[Intemann et al., 2023].

5.3. Ausblick

In dieser Arbeit wurden Konfigurationen der bestehenden Record Linkage Lösungen E-PIX und FRIL erarbeitet, die erfolgreich einen Großteil der Kriterien zur Reduzierung von Verknüpfungsfehlern erfüllen. Es bleibt jedoch eine offene Frage, inwiefern andere Record Linkage Lösungen die vorliegenden Kriterien in ihrer Konfigurierbarkeit berücksichtigen können und in welchen Bereichen Optimierungsbedarf besteht. Die Beantwortung dieser Frage erfordert zusätzliche Tests mit verschiedenen Record Linkage Lösungen, die über den Umfang dieser Arbeit hinausgehen. In solchen Studien könnte ein Goldstandard-Datensatz als Referenz dienen, der alle bekannten Fehlerfälle enthält.
Angesichts der Erkenntnis, dass der Record Linkage Prozess in der Forschung eine entscheidende Rolle in Bezug auf die Forschungsergebnisse spielt und Verknüpfungsfehler schwerwiegende Folgen für die Forschung und die menschliche Gesundheit haben können, entsteht ein weiterer Forschungsbedarf. Es könnte von Nutzen sein, weitere Untersuchungen zur Minimierung von Synonymfehlern durchzuführen, um die Genauigkeit und Zuverlässigkeit von Studien zu verbessern. Hieraus ergeben sich mögliche Untersuchungen, die beispielsweise die Verwendung von maschinellen Lernansätzen im Verknüpfungsprozess analysieren und ihren Einfluss auf die Minimierung von Verknüpfungsfehlern untersuchen.
Die in dieser Arbeit entwickelte Empfehlung für die Nutzung einer geeigneten Konfiguration in Registern könnte die Verknüpfungsqualität des Record Linkage in diesem Bereich in Zukunft erhöhen. Allerdings ist zu beachten, dass Record Linkage in vielen weiteren Anwendungsbereichen zum Einsatz kommt, für die solche Konfigurationen nicht existieren oder nicht frei zugänglich sind. Daher könnte es lohnend sein, Empfehlungen für geeignete Konfigurationen in verschiedenen Bereichen zu entwickeln, um die Qualität der Verknüpfungs zu verbessern und die Folgen dieser Fehler zu vermeiden.
Die dargestellte Problematik eines derzeit fehlenden einheitlichen Goldstandard-

Datensatzes behindert die Optimierung und den Vergleich von bestehenden Record Linkage Lösungen. Diese Erkenntnis sollte Anlass sein, Möglichkeiten zu entwickeln, einen Goldstandard zu erstellen, mit Hilfe dessen eine Analyse der Verknüpfungsqualität in allen Record Linkage Lösungen erfolgen kann. In diesem Zusammenhang steht die Problematik des Zugriffs auf Echtdaten. Es bleibt daher abzuwarten, ob die Entwicklung eines solchen Goldstandards in Zukunft umgesetzt werden kann.
Aufgrund der gegenwärtigen Herausforderung, einen eindeutigen Identifikator für die Verknüpfung von Patientendaten zu erhalten, wird in der Anwendung vermutlich der probabilistische Verknüpfungsansatz weiterhin einem deterministischen Ansatz vorgezogen werden. Die Entwicklung eines solchen Identifikators würde das Record Linkage in Zukunft deutlich verbessern und die Verknüpfungsqualität massiv erhöhen. Doch aufgrund der genannten Herausforderungen gestaltet sich eine Entwicklung schwierig und bleibt vorerst unklar.

Literaturverzeichnis

[Adelaide et al., 2014] Adelaide, A., Bakker, B., de Groot, M., Grootheest, G., van der Laan, J., Smit, J., and Verkerk, B. (2014). Record linkage in health data. *CBS, Den Haag*, pages 3–64.

[Bellow et al., 2016] Bellow, M. E., Daniel, K., Gorsak, M., and Erciulescu, A. L. (2016). Evaluating Record Linkage Software for Agricultural Surveys. *JSM Proceedings. Survey Research Methods Section. Alexandria, VA: American Statistical Association.*, pages 3225–3235.

[Bergman et al., 2000] Bergman, L., Beelen, M. L., Gallee, M. P., Hollema, H., Benraadt, J., and van Leeuwen, F. E. (2000). Risk and prognosis of endometrial cancer after tamoxifen for breast cancer. *The Lancet*, 356(9233):881–887.

[Bialke et al., 2015a] Bialke, M., Bahls, T., Havemann, C., Piegsa, J., Weitmann, K., Wegner, T., and Hoffmann, W. (2015a). MOSAIC – A Modular Approach to Data Management in Epidemiological Studies. *Methods of Information in Medicine*, 54(04):364–371.

[Bialke et al., 2015b] Bialke, M., Penndorf, P., Wegner, T., Bahls, T., Havemann, C., Piegsa, J., and Hoffmann, W. (2015b). A workflow-driven approach to integrate generic software modules in a Trusted Third Party. *Journal of Translational Medicine*, 13(1):1–8.

[Blakely and Salmond, 2002] Blakely, T. and Salmond, C. (2002). Probabilistic record linkage and a method to calculate the positive predictive value. *International Journal of Epidemiology*, 31(6):1246–1252.

[Boettcher et al., 2014] Boettcher, S., Hartel, R., and Hawicks, H. (2014). Namenlemmatisierung in der web-datenbank mittelalterlicher und frühneuzeitlicher universitätsmatrikel. *Els noms en la vida quotidiana. Actes del XXIV Congrés Internacional d'ICOS sobre Ciències Onomàstiques*, pages 283–293.

[Bozkurt et al., 2009] Bozkurt, O., de Boer, A., Grobbee, D. E., de Leeuw, P. W., Kroon, A. A., Schiffers, P., and Klungel, O. H. (2009). Variation in Renin-Angiotensin System and Salt-Sensitivity Genes and the Risk of Diabetes Mellitus Associated With the Use of Thiazide Diuretics. *American Journal of Hypertension*, 22(5):545–551.

[Campbell, 2005] Campbell, K. (2005). Rule Your Data with The Link King (a SAS/AF application for record linkage and unduplication), in SUGI 30 Proceedings. *SUGI 30*, pages 1–9.

[ChoiceMaker, 2023] ChoiceMaker(URL:https://choicemaker.com/(abgerufen: 16.11.2023 06:03, 2023,)). ChoiceMaker. *Website*.

[Christen, 2008] Christen, P. (2008). Febrl: a freely available record linkage system with a graphical user interfac. In *the second Australasian workshop on Health data and knowledge management*, volume 80, pages 14–25.

[da Silveira and Artmann, 2009] da Silveira, D. P. and Artmann, E. (2009). Acurácia em métodos de relacionamento probabilístico de bases de dados em saúde: revisão sistemática. *Revista de Saúde Pública*, 43(5):875–882.

[DataMatch, 2023] DataMatch (URL:https://dataladder.com/de/produkte/datamatch-enterprise-bewertetes-datenqualitaetsmanagementprodukt-nr-1/(abgerufen: 17.11.2023 06:20, 2023,)). DataMatch. *Website*.

[Derczynski, 2016] Derczynski, L. (2016). Complementarity, F-score, and NLP evaluation. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 261–266, Portorož, Slovenia. European Language Resources Association (ELRA).

[Doidge et al., 2020] Doidge, J., Christen, P., and Harron, K. (2020). Quality assessment in data linkage. In *Joined up data in government: the future of data linking methods [Internet].*, pages 1–16. UK Government Analysis Function and Office for National Statistics.

[Doidge and Harron, 2019] Doidge, J. C. and Harron, K. L. (2019). Reflections on modern methods: linkage error bias. *International Journal of Epidemiology*, pages 2050–2060.

[DSVGO, 2024] DSVGO(URL:https://dsgvo-gesetz.de/art-5-dsgvo/(abgerufen: 05.01.2024 18:16, 2024,)). Art.5 DSGVO Grundsätze für die Verarbeitung personenbezogener Daten, c. *Website*.

[Dusetzina SB, 2014] Dusetzina SB, Tyree S, M. A. M. A. G. L. C. W. (2014). Linking Data for Health Services Research: A Framework and Instructional Guide. *Linking Data for Health Services Research: A Framework and Instructional Guide [Internet]*, pages 1–77.

[Fair, 2004] Fair, M. (2004). Generalized record linkage system–Statistics Canada's record linkage software. *Austrian Journal of Statistics*, 33(1&2):37–53.

[Fellegi and Sunter, 1969] Fellegi, I. P. and Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, 64(328):1183–1210.

[Franke et al., 2019] Franke, M., Sehili, Z., and Rahm, E. (2019). PRIMAT. *Proceedings of the VLDB Endowment*, 12(12):1826–1829.

[FRIL, 2023] FRIL (URL:https://fril.sourceforge.net/(abgerufen: 17.11.2023 16:03, 2023,)). FRIL. *Website*.

[Hampf, 2021] Hampf, C. (2021). Anwenderhandbuch E-PIX, Version 2.10. *Greifswald*, pages 1–45.

[Hampf et al., 2020] Hampf, C., Geidel, L., Zerbe, N., Bialke, M., Stahl, D., Blumentritt, A., Bahls, T., Hufnagl, P., and Hoffmann, W. (2020). Assessment of scalability and performance of the record linkage tool E-PIX® in managing multi-million patients in research projects at a large university hospital in Germany. *Journal of Translational Medicine*, 18(1):1–11.

[Harron, 2022] Harron, K. (2022). Data linkage in medical research. *BMJ Medicine*, 1(1):1–3.

[Harron et al., 2017a] Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M. L., and Goldstein, H. (2017a). Challenges in administrative data linkage for research. *Big Data &amp Society*, 4(2):1–12.

[Harron et al., 2020] Harron, K., Doidge, J. C., and Goldstein, H. (2020). Assessing data linkage quality in cohort studies. *Annals of Human Biology*, 47(2):218–226.

[Harron et al., 2017b] Harron, K. L., Doidge, J. C., Knight, H. E., Gilbert, R. E., Goldstein, H., Cromwell, D. A., and van der Meulen, J. H. (2017b). A guide to evaluating linkage quality for the analysis of linked data. *International Journal of Epidemiology*, 46(5):1699–1710.

[Intemann et al., 2023] Intemann, T., Kaulke, K., Kipker, D.-K., Lettieri, V., Stallmann, C., Schmidt, C. O., Geidel, L., Bialke, M., Hampf, C., Stahl, D., et al. (2023). White Paper - Verbesserung des Record Linkage für die Gesundheitsforschung in Deutschland : August 2023. *arXiv preprint arXiv:2312.10093*, pages 1–167.

[Jurczyk, 2009] Jurczyk, P. (2009). Fine-grained Record Integration and Linkage Tool Tutorial V3.2. *Atlanta, Georgia*, pages 1–43.

[Jurczyk et al., 2008] Jurczyk, P., Lu, J., Xiong, L., Cragan, J., and Correa, A. (2008). FRIL: A tool for comparative record linkage. *AMIA Annual Symposium proceedings*, 2008:440–444.

[Kvalsvig et al., 2019] Kvalsvig, A., Gibb, S., and Teng, A. (2019). Linkage error and linkage bias: A guide for IDI users. *University of Otago*, pages 1–32.

[Kötzschke, 2015] Kötzschke, G. (Stralsund, 2015). Bewertung von Identitätsmanagement-Tools anhand von Anwendungsfällen klinischepidemiologischer Forschungsprojekte einschließlich der Erarbeitung ausgewählter Testszenarien. *Master-Thesis, Fachhochschule Stralsund*.

[LinkageWiz, 2023] LinkageWiz (URL:http://www.linkagewiz.net/index.htm (abgerufen: 16.11.2023 15:40, 2023,)). Linkagewiz. *Website*.

[LinkKing, 2023] LinkKing(URL:http://the-link-king.party/(abgerufen:17.11.2023 12:46, 2023,)). Linkking. *Website*.

[Lisbach, 2011] Lisbach, B. (Deutschland, 2011). Einleitung: Paradigmenwechsel im Identity Matching. In *Linguistisches Identity Matching*, pages 1–88. ViewegTeubner.

[Mainzelliste, 2023] Mainzelliste(URL:https://www.unimedizin-mainz.de/imbei/informatik/ag-verbundforschung/mainzelliste.html(abgerufen: 06.01.2024 15:00, 2023,)). Mainzellitse. *Website, Universitätsmedizin Mainz*.

[March et al., 2019] March, S., Andrich, S., Drepper, J., Horenkamp-Sonntag, D., Icks, A., Ihle, P., Kieschke, J., Kollhorst, B., Maier, B., Meyer, I., Müller, G., Ohlmeier, C., Peschke, D., Richter, A., Rosenbusch, M.-L., Scholten, N., Schulz, M., Stallmann, C., Swart, E., Wobbe-Ribinski, S., Wolter, A., Zeidler, J., and Hoffmann, F. (2019). Gute Praxis Datenlinkage (GPD). *Das Gesundheitswesen*, 81(08/09):636–650.

[March et al., 2018] March, S., Antoni, M., Kieschke, J., Kollhorst, B., Maier, B., Müller, G., Sariyar, M., Schulz, M., Enno, S., Zeidler, J., and Hoffmann, F. (2018). Quo vadis Datenlinkage in Deutschland? Eine erste Bestandsaufnahme. *Das Gesundheitswesen*, 57(03):e20–e31.

[Moore et al., 2014] Moore, C. L., Amin, J., Gidding, H. F., and Law, M. G. (2014). A New Method for Assessing How Sensitivity and Specificity of Linkage Studies Affects Estimation. *PLoS ONE*, 9(7):1–6.

[OpenEMPI, 2023a] OpenEMPI(URL:https://www.openempi.org/(abgerufen: 17.11.2023 16:03, 2023,)a). Openempi. *Website*.

[OpenEMPI, 2023b] OpenEMPI(URL:https://openempi.atlassian.net/wiki/spaces/openempi30/pages/1174503429/Settings Page(abgerufen: 18.11.2023 06:24, 2023,)b). OpenEMPISettings. *Website*.

[Primat, 2023] Primat (URL:https://www.toolpool-gesundheitsforschung.de/produkte/primat-private-matching-toolbox (abgerufen: 16.11.2023 12:34, 2023,)). Primat. *Website*.

[Rau et al., 2020] Rau, H., Geidel, L., Hampf, C., Bahls, T., Stahl, D., Blumentritt, A., and Bialke, M. (2020). Trusted Third Party solutions - Demonstration of 3 Software Tools for EU-GDPR compliant Record Linkage, Consent Management and Pseudonymisation. *Medical Informatics Europe conference (MIE)*, pages 1–3.

[Rohde et al., 2021] Rohde, F., Franke, M., Sehili, Z., Lablans, M., and Rahm, E. (2021). Optimization of the Mainzelliste software for fast privacy-preserving record linkage. *Journal of Translational Medicine*, 19:1–12.

[Sariyar et al., 2011] Sariyar, M., Borg, A., and Pommerening, K. (2011). Controlling false match rates in record linkage using extreme value theory. *Journal of Biomedical Informatics*, 44(4):648–654.

[Sayers et al., 2015] Sayers, A., Ben-Shlomo, Y., Blom, A. W., and Steele, F. (2015). Probabilistic record linkage. *International Journal of Epidemiology*, 45(3):954–964.

[Schmidtmann et al., 2016] Schmidtmann, I., Sariyar, M., Borg, A., Gerold-Ay, A., Heidinger, O., Hense, H.-W., Krieg, V., and Hammer, G. P. (2016). Quality of record linkage in a highly automated cancer registry that relies on encrypted identity data. *GMS Medizinische Informatik*, pages 1–11.

[Schuster, 2002] Schuster, J. (2002). Rechtschreibkorrektur-Probabilistic models of pronunciation and spelling -. *Vortrag im Rahmen des Proseminars Computerlinguistik 1 an der LMU München. URL:https://www.cis.uni-muenchen.de/ micha/praesentationen /rechtschreibkorrektur/ArtenRechtschreibfehler.html(abgerufen:06.01.2024,08:28)*.

[THS, 2023b] THS (URL: https://www.ths-greifswald.de/ (abgerufen: 07.10.2023 16:03, 2023b). Unabhängige Treuhandstelle Universitätsmedizin Greifswald. *Website*.

[THS, 2023a] THS (URL:https://www.ths-greifswald.de/forscher/e-pix/dokumentation (abgerufen: 15.11.2023 18:15, 2023,)a). E-PIX, Dokumentation. *Website*.

[THS-Community, 2024] THS-Community (URL: https://github.com/ths-community (abgerufen: 19.02.2023 13:37, 2024,)). Record-Linkage Konfiguration des E-PIX als Grundlage für Register. *github*.

[THS, 2022] THS(Record Linkage und Identitätsmanagement, 2022). URL:https://.ths-greifswald.de/wp-content/uploads/2022/12/e-pix$_B$roschuere$_V$06$_S$tand$_1$−Dez.pdf.Broschüre, UnabhängigeTreuhandstelleGreifswald.

[THS, 2023] THS(URL:https://www.ths-greifswald.de/forscher/e-pixverbreitung(abgerufen: 15.11.2023 17:50, 2023,)). E-pix, verbreitung. *Website*.

[Tromp et al., 2006] Tromp, M., Reitsma, J., Ravelli, A. C. J., Méray, N., and Bonsel, G. (2006). Record linkage: making the most out of errors in linking variables. *AMIA Annual Symposium proceedings*, page 779—783.

[Vatsalan et al., 2017] Vatsalan, D., Sehili, Z., Christen, P., and Rahm, E. (2017). Privacy-preserving record linkage for big data: Current approaches and research challenges. pages 851–895.

[Weiand, 2022] Weiand, S. (2022). Vergleich von Record-Linkage Methoden anhand der Mikro-Simulation eines bundesweiten Schülerregisters. *GRLC Working Paper Series*, pages 1–84.

[Winkler, 2000] Winkler, W. E. (2000). Frequency-based matching in fellegi-sunter model of record linkage. *Bureau of the Census Statistical Research Division*, pages 1–13.

[Winkler, 2003] Winkler, W. E. (Washington, DC, 2003). Data cleaning methods. In *Proceedings of the KDD-2003 Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, pages 1–6.

[Zhu et al., 2009] Zhu, V. J., Overhage, M. J., Egg, J., Downs, S. M., and Grannis, S. J. (2009). An Empiric Modification to the Probabilistic Record Linkage Algorithm Using Frequency-Based Weight Scaling. *Journal of the American Medical Informatics Association*, 16(5):738–745.

[Zhu et al., 2015] Zhu, Y., Matsuyama, Y., Ohashi, Y., and Setoguchi, S. (2015). When to conduct probabilistic linkage vs. deterministic linkage? A simulation study. *Journal of Biomedical Informatics*, 56:80–86.

A. Record Linkage Lösungen

	Kostenfrei	Unterstützte Record-Linkage Verfahren	Schnittstellen	Unterstützte Ähnlichkeitsmaße	Grafische Benutzeroberfläche	Datentransformation	Nutzung künstlicher Intelligenz	Verwendung in der Forschung	Entwickler	Suchmethode
E-PIX	Ja	Probabilistisches & deterministisches Linkage, PPRL	SOAP-Webschnittstelle	Levenshtein Algorithmus, phonetische & deterministische Algorithmen	Ja	Ja	Nein	Ja	Treuhandstelle der Universitätsmedizin Greifswald	Blocking
ChoiceMaker	Ja	Probabilistisches & deterministisches Linkage	SOAP-Webschnittstelle	Soundex, Edit-Distance, Jaro-Winkler, NYSIIS, Metaphone, Double-Metaphone, Value-Frequency-Weighting , Levenshtein-Distanz	Ja	Ja	Ja (Entwicklung von Record-Matching Modellen)	Ja	ChoiceMaker Technologies (New Jersey, USA)	Blocking
Mainzelliste	Ja	EpiLink-Algorithmus, deterministisches Linkage, PPRL mit Bloomfilter & Secure Multi Party Computation	REST-basierte Webschnittstelle	Dice-Ähnlichkeit, Binäre-Ähnlichkeit, nGramm Methode, Wertegleichheit	Ja	Ja	Nein	Ja	Mainzelliste Community (IMBEI der Universität Mainz)	Blocking
Primat	Ja	Probabilistisches & deterministisches Linkage, PPRL mit Bloomfilter & Härtungstechniken	n.A.	Jaccard-Ähnlichkeit, Dice-Ähnlichkeit, Hamming-Ähnlichkeit	Nein	Ja	Nein	Nein (noch in der Entwicklung)	Universität Leipzig	Blocking
LinkKing	Ja (2000$ Lizenz für Base SAS)	Probabilistisches (Algorithmen von MEDSTAT) & deterministisches Linkage	Benutzeroberfläche	Soundex Funktion von SAS, Approximate String Matching-Algorithmus, phonetische Äquivalenzfunktion des New York State Intelligence Information System	Ja	Ja	Ja (Datenaufbereitung und Verknüpfung optimieren)	n.A.	Abteilung für Alkohol und Drogenmissbrauch des Bundesstaates Washington	Blocking
FRIL	Ja	Probabilistisches & deterministisches Linkage	Benutzeroberfläche	Edit Distance, Soundex, n-Gramm , Gleichheit	Ja	Ja	Nein (Einsatz in Planung)	Ja	Zusammenarbeit: Emory University & Centers for Disease Control and Prevention (Atlanta, Georgia)	Sortierte Nachbarschaftsmethode, Blocking
Febrl	Ja	Probabilistisches & deterministisches Linkage	Benutzeroberfläche	26 verschiedene Vergleichsmethoden[1] (Approximation-String Vergleiche & numerische Vergleich)	Ja	Ja	Nein	Ja	Australian National University Data Mining Group (Canberra, Australien)	Qgramindex, Fuzzy-Blocking, Canopyindex, StringMapIndex, SuffixArrayIndex
OpenEMPI	Ja	Probabilistisches & deterministisches Linkage	IHE PIX/PDQ- & Rest-basierte Webschnittstelle	Deterministischer Algorithmus, Integration von probabilistischen Maßen möglich. z.B Levenshtein-Distanz & Jaro-Winklern	Ja	Ja	Ja Version 4.3.0 (Ähnlichkeitsmetrik und Schwellenwert bestimmen)	Ja	SYSNET International	Blocking, SuffixArrayIndex, Sortierte Nachbschaftsmethode
G-Link	Nein (12.500CS)	Probabilistisches Linkage	Benutzeroberfläche	Deterministische Algorithmen, NYSIIS, Tippfehler-Vereinbarung	Ja	Nein	Nein	n.A.	Statistics Canada	Blocking
LinkageWiz	Nein (2.999$)	Probabilistisches & deterministisches Linkage	Benutzeroberfläche	Phonetisch Algorithmen: NYSIIS & SOUNDEX ; String-Vergleichsfunktionen: Levenshtein-Distanz	Ja	Ja	Nein	Ja	LinkageWiz Software (Australien)	Blocking
DataMatch	Nein (keine Preisangabe)	Probabilistisches & deterministisches Linkage	RESTful-API	Levenshtein-Distanz, Damerau-Levenshtein-Abstand, Jaro-Winkler Abstand,Tastaturabstand, Kullback Leibler-Abstand, Jaccard-Index, Metaphon 3, Name Variante, Silbenausrichtung, Akronym	Ja	Ja	Nein	Ja	DataLadder	Blocking

[1][Christen, 2008] Christen, P. (2008), [S.5] Febrl: a freely available record linkage system with a graphical user interfac. in the second Australasian workshop on Healt

B. Konfigurationen des E-PIX

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<ns2:MatchingConfiguration xmlns:ns2="http://www.ttp.icmvc.emau.org/deduplication/config/model">
    <matching-mode>MATCHING_IDENTITIES</matching-mode>
    <mpi-generator>org.emau.icmvc.ttp.epix.gen.impl.EAN13Generator</mpi-generator>
    <mpi-prefix>1001</mpi-prefix>
    <use-notifications>false</use-notifications>
    <limit-search-to-reduce-memory-consumption>false</limit-search-to-reduce-memory-consumption>
    <persist-mode>IDENTIFYING</persist-mode>
    <required-fields>
        <name>firstName</name>
        <name>lastName</name>
        <name>birthDate</name>
        <name>gender</name>
    </required-fields>
    <preprocessing-config>
        <preprocessing-field>
            <field-name>firstName</field-name>
            <simple-transformation-type>
                <input-pattern>Dr.</input-pattern>
                <output-pattern></output-pattern>
            </simple-transformation-type>
            <simple-transformation-type>
                <input-pattern>med.</input-pattern>
                <output-pattern></output-pattern>
            </simple-transformation-type>
            <simple-transformation-type>
                <input-pattern>Dipl.</input-pattern>
                <output-pattern></output-pattern>
            </simple-transformation-type>
            <simple-transformation-type>
                <input-pattern>          </input-pattern>
                <output-pattern>  </output-pattern>
```

```
33              </simple-transformation-type>
34              <simple-transformation-type>
35                  <input-pattern>Prof.</input-pattern>
36                  <output-pattern></output-pattern>
37              </simple-transformation-type>
38              <simple-transformation-type>
39                  <input-pattern>rer.</input-pattern>
40                  <output-pattern></output-pattern>
41              </simple-transformation-type>
42              <simple-transformation-type>
43                  <input-pattern>,</input-pattern>
44                  <output-pattern></output-pattern>
45              </simple-transformation-type>
46              <simple-transformation-type>
47                  <input-pattern>-</input-pattern>
48                  <output-pattern></output-pattern>
49              </simple-transformation-type>
50              <simple-transformation-type>
51                  <input-pattern>nat.</input-pattern>
52                  <output-pattern></output-pattern>
53              </simple-transformation-type>
54              <simple-transformation-type>
55                  <input-pattern>Ing.</input-pattern>
56                  <output-pattern></output-pattern>
57              </simple-transformation-type>
58              <simple-transformation-type>
59                  <input-pattern>?</input-pattern>
60                  <output-pattern></output-pattern>
61              </simple-transformation-type>
62              <complex-transformation-type>
63                  <qualified-class-name>org.emau.icmvc.ttp.
                        deduplication.preprocessing.impl.
                        ToUpperCaseTransformation</qualified-
                        class-name>
64              </complex-transformation-type>
65              <complex-transformation-type>
66                  <qualified-class-name>org.emau.icmvc.ttp.
                        deduplication.preprocessing.impl.
                        CharsMutationTransformation</qualified-
                        class-name>
67              </complex-transformation-type>
68          </preprocessing-field>
69          <preprocessing-field>
```

```
70              <field-name>lastName</field-name>
71              <simple-transformation-type>
72                  <input-pattern>Dr.</input-pattern>
73                  <output-pattern></output-pattern>
74              </simple-transformation-type>
75              <simple-transformation-type>
76                  <input-pattern>Dipl.</input-pattern>
77                  <output-pattern></output-pattern>
78              </simple-transformation-type>
79              <simple-transformation-type>
80                  <input-pattern>         </input-pattern>
81                  <output-pattern>  </output-pattern>
82              </simple-transformation-type>
83              <simple-transformation-type>
84                  <input-pattern>Prof.</input-pattern>
85                  <output-pattern></output-pattern>
86              </simple-transformation-type>
87              <simple-transformation-type>
88                  <input-pattern>,</input-pattern>
89                  <output-pattern></output-pattern>
90              </simple-transformation-type>
91              <simple-transformation-type>
92                  <input-pattern>-</input-pattern>
93                  <output-pattern></output-pattern>
94              </simple-transformation-type>
95              <simple-transformation-type>
96                  <input-pattern>Ing.</input-pattern>
97                  <output-pattern></output-pattern>
98              </simple-transformation-type>
99              <simple-transformation-type>
100                 <input-pattern>med.</input-pattern>
101                 <output-pattern></output-pattern>
102             </simple-transformation-type>
103             <simple-transformation-type>
104                 <input-pattern>rer.</input-pattern>
105                 <output-pattern></output-pattern>
106             </simple-transformation-type>
107             <simple-transformation-type>
108                 <input-pattern>
109         </input-pattern>
110                 <output-pattern></output-pattern>
111             </simple-transformation-type>
112             <simple-transformation-type>
```

```
113                 <input-pattern>nat.</input-pattern>
114                 <output-pattern></output-pattern>
115             </simple-transformation-type>
116             <simple-transformation-type>
117                 <input-pattern>?</input-pattern>
118                 <output-pattern></output-pattern>
119             </simple-transformation-type>
120             <complex-transformation-type>
121                 <qualified-class-name>org.emau.icmvc.ttp.
                        deduplication.preprocessing.impl.
                        ToUpperCaseTransformation</qualified-
                        class-name>
122             </complex-transformation-type>
123             <complex-transformation-type>
124                 <qualified-class-name>org.emau.icmvc.ttp.
                        deduplication.preprocessing.impl.
                        CharsMutationTransformation</qualified-
                        class-name>
125             </complex-transformation-type>
126         </preprocessing-field>
127     </preprocessing-config>
128     <matching>
129         <threshold-possible-match>12.99</threshold-possible-
                match>
130         <threshold-automatic-match>14.5</threshold-automatic
                -match>
131         <use-cemfim>false</use-cemfim>
132         <parallel-matching-after>1000</parallel-matching-
                after>
133         <number-of-threads-for-matching>4</number-of-threads
                -for-matching>
134         <field>
135             <name>firstName</name>
136             <blocking-threshold>0.4</blocking-threshold>
137             <blocking-mode>TEXT</blocking-mode>
138             <matching-threshold>1</matching-threshold>
139             <weight>8</weight>
140             <algorithm>org.emau.icmvc.ttp.deduplication.impl
                    .LevenshteinAlgorithm</algorithm>
141             <multiple-values>
142                 <separator> </separator>
143                 <penalty-not-a-perfect-match>0.5</penalty-
                        not-a-perfect-match>
```

```
144                 <penalty-one-short>0.5</penalty-one-short>
145                 <penalty-both-short>0.5</penalty-both-short>
146             </multiple-values>
147         </field>
148         <field>
149             <name>lastName</name>
150             <matching-threshold>1</matching-threshold>
151             <weight>6</weight>
152             <algorithm>org.emau.icmvc.ttp.deduplication.impl
                    .LevenshteinAlgorithm</algorithm>
153         </field>
154         <field>
155             <name>gender</name>
156             <matching-threshold>0.75</matching-threshold>
157             <weight>3</weight>
158             <algorithm>org.emau.icmvc.ttp.deduplication.impl
                    .LevenshteinAlgorithm</algorithm>
159         </field>
160         <field>
161             <name>birthDate</name>
162             <blocking-threshold>0.6</blocking-threshold>
163             <blocking-mode>NUMBERS</blocking-mode>
164             <matching-threshold>1</matching-threshold>
165             <weight>9</weight>
166             <algorithm>org.emau.icmvc.ttp.deduplication.impl
                    .LevenshteinAlgorithm</algorithm>
167         </field>
168     </matching>
169 </ns2:MatchingConfiguration>
```

Listing B.1: Konfiguration 1

```
<matching>
    <threshold-possible-match>25.99</threshold-possible-
       match>
    <threshold-automatic-match>29.5</threshold-automatic-
       match>
    <use-cemfim>false</use-cemfim>
    <parallel-matching-after>1000</parallel-matching-after>
    <number-of-threads-for-matching>4</number-of-threads-for
       -matching>
    <field>
      <name>firstName</name>
      <blocking-threshold>0.7</blocking-threshold>
      <matching-threshold>0.8</matching-threshold>
      <weight>6</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         ColognePhoneticAlgorithm</algorithm>
      <multiple-values>
        <separator> </separator>
        <penalty-not-a-perfect-match>0.2</penalty-not-a-
           perfect-match>
        <penalty-one-short>0.6</penalty-one-short>
        <penalty-both-short>0.6</penalty-both-short>
      </multiple-values>
    </field>
    <field>
      <name>lastName</name>
      <matching-threshold>0.8</matching-threshold>
      <weight>4</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         ColognePhoneticAlgorithm</algorithm>
    </field>
    <field>
      <name>gender</name>
      <matching-threshold>0.75</matching-threshold>
      <weight>6</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         ColognePhoneticAlgorithm</algorithm>
    </field>
    <field>
      <name>birthDate</name>
      <blocking-threshold>0.9</blocking-threshold>
      <blocking-mode>NUMBERS</blocking-mode>
      <matching-threshold>1</matching-threshold>
```

```
37        <weight>14</weight>
38        <algorithm>org.emau.icmvc.ttp.deduplication.impl.
              CologneAlgorithm</algorithm>
39      </field>
40    </matching>
```

Listing B.2: Konfiguration 2

```
<matching>
    <threshold-possible-match>12.99</threshold-possible-
       match>
    <threshold-automatic-match>14.5</threshold-automatic-
       match>
    <use-cemfim>false</use-cemfim>
    <parallel-matching-after>1000</parallel-matching-after>
    <number-of-threads-for-matching>4</number-of-threads-for
       -matching>
    <field>
      <name>firstName</name>
      <blocking-threshold>1</blocking-threshold>
      <matching-threshold>1</matching-threshold>
      <weight>8</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         DeterministicAlgorithm</algorithm>
      <multiple-values>
        <separator> </separator>
        <penalty-not-a-perfect-match>0.5</penalty-not-a-
           perfect-match>
        <penalty-one-short>0.5</penalty-one-short>
        <penalty-both-short>0.5</penalty-both-short>
      </multiple-values>
    </field>
    <field>
      <name>lastName</name>
      <matching-threshold>1</matching-threshold>
      <weight>6</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         DeterministicAlgorithm</algorithm>
    </field>
    <field>
      <name>gender</name>
      <matching-threshold>1</matching-threshold>
      <weight>12</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         DeterministicAlgorithm</algorithm>
    </field>
    <field>
      <name>birthDate</name>
      <blocking-threshold>1</blocking-threshold>
      <blocking-mode>NUMBERS</blocking-mode>
      <matching-threshold>1</matching-threshold>
```

```
37        <weight>14</weight>
38        <algorithm>org.emau.icmvc.ttp.deduplication.impl.
              DeterministicAlgorithm</algorithm>
39      </field>
40    </matching>
```

Listing B.3: Konfiguration 3

```
<matching>
    <threshold-possible-match>12.99</threshold-possible-
       match>
    <threshold-automatic-match>14.5</threshold-automatic-
       match>
    <use-cemfim>false</use-cemfim>
    <parallel-matching-after>1000</parallel-matching-after>
    <number-of-threads-for-matching>4</number-of-threads-for
       -matching>
    <field>
      <name>firstName</name>
      <blocking-threshold>0.4</blocking-threshold>
      <matching-threshold>0.8</matching-threshold>
      <weight>8</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         LevenshteinAlgorithm</algorithm>
      <multiple-values>
        <separator> </separator>
        <penalty-not-a-perfect-match>0.1</penalty-not-a-
           perfect-match>
        <penalty-one-short>0.2</penalty-one-short>
        <penalty-both-short>0.2</penalty-both-short>
      </multiple-values>
    </field>
    <field>
      <name>lastName</name>
      <matching-threshold>0.8</matching-threshold>
      <weight>6</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         LevenshteinAlgorithm</algorithm>
    </field>
    <field>
      <name>gender</name>
      <matching-threshold>0.75</matching-threshold>
      <weight>3</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         LevenshteinAlgorithm</algorithm>
    </field>
    <field>
      <name>birthDate</name>
      <blocking-threshold>0.6</blocking-threshold>
      <blocking-mode>NUMBERS</blocking-mode>
      <matching-threshold>1</matching-threshold>
```

```
37        <weight>9</weight>
38        <algorithm>org.emau.icmvc.ttp.deduplication.impl.
              LevenshteinAlgorithm</algorithm>
39      </field>
40    </matching>
```

Listing B.4: Konfiguration 4

```
<matching>
    <threshold-possible-match>15.99</threshold-possible-
       match>
    <threshold-automatic-match>17.5</threshold-automatic-
       match>
    <use-cemfim>false</use-cemfim>
    <parallel-matching-after>1000</parallel-matching-after>
    <number-of-threads-for-matching>4</number-of-threads-for
       -matching>
    <field>
      <name>firstName</name>
      <blocking-threshold>0.7</blocking-threshold>
      <matching-threshold>0.8</matching-threshold>
      <weight>9</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         ColognePhoneticAlgorithm</algorithm>
      <multiple-values>
        <separator> </separator>
        <penalty-not-a-perfect-match>0.2</penalty-not-a-
           perfect-match>
        <penalty-one-short>0.6</penalty-one-short>
        <penalty-both-short>0.6</penalty-both-short>
      </multiple-values>
    </field>
    <field>
      <name>lastName</name>
      <matching-threshold>0.8</matching-threshold>
      <weight>5</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         ColognePhoneticAlgorithm</algorithm>
    </field>
    <field>
      <name>gender</name>
      <matching-threshold>0.75</matching-threshold>
      <weight>6</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         ColognePhoneticAlgorithm</algorithm>
    </field>
    <field>
      <name>birthDate</name>
      <blocking-threshold>0.9</blocking-threshold>
      <blocking-mode>NUMBERS</blocking-mode>
      <matching-threshold>1</matching-threshold>
```

```
37        <weight>14</weight>
38        <algorithm>org.emau.icmvc.ttp.deduplication.impl.
              CologneAlgorithm</algorithm>
39      </field>
40      <field>
41        <name>value3</name>
42        <matching-threshold>0.75</matching-threshold>
43        <weight>9</weight>
44        <algorithm>org.emau.icmvc.ttp.deduplication.impl.
              CologneAlgorithm</algorithm>
45      </field>
46    </matching>
```

Listing B.5: Konfiguration 5

```
<matching>
    <threshold-possible-match>12.99</threshold-possible-
       match>
    <threshold-automatic-match>14.5</threshold-automatic-
       match>
    <use-cemfim>false</use-cemfim>
    <parallel-matching-after>1000</parallel-matching-after>
    <number-of-threads-for-matching>4</number-of-threads-for
       -matching>
    <field>
      <name>firstName</name>
      <blocking-threshold>1</blocking-threshold>
      <matching-threshold>1</matching-threshold>
      <weight>8</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         DeterministicAlgorithm</algorithm>
      <multiple-values>
        <separator> </separator>
        <penalty-not-a-perfect-match>0.5</penalty-not-a-
           perfect-match>
        <penalty-one-short>0.5</penalty-one-short>
        <penalty-both-short>0.5</penalty-both-short>
      </multiple-values>
    </field>
    <field>
      <name>lastName</name>
      <matching-threshold>1</matching-threshold>
      <weight>6</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         DeterministicAlgorithm</algorithm>
    </field>
    <field>
      <name>gender</name>
      <matching-threshold>1</matching-threshold>
      <weight>12</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         DeterministicAlgorithm</algorithm>
    </field>
    <field>
      <name>birthDate</name>
      <blocking-threshold>1</blocking-threshold>
      <blocking-mode>NUMBERS</blocking-mode>
      <matching-threshold>1</matching-threshold>
```

```
37        <weight>14</weight>
38        <algorithm>org.emau.icmvc.ttp.deduplication.impl.
              DeterministicAlgorithm</algorithm>
39      </field>
40      <field>
41        <name>value3</name>
42        <matching-threshold>1</matching-threshold>
43        <weight>9</weight>
44        <algorithm>org.emau.icmvc.ttp.deduplication.impl.
              DeterministicAlgorithm</algorithm>
45      </field>
46    </matching>
```

Listing B.6: Konfiguration 6

```
<matching>
  <threshold-possible-match>10.5</threshold-possible-match>
  <threshold-automatic-match>14.5</threshold-automatic-match>
  <use-cemfim>false</use-cemfim>
  <parallel-matching-after>1000</parallel-matching-after>
  <number-of-threads-for-matching>4</number-of-threads-for-matching>
  <field>
    <name>firstName</name>
    <blocking-threshold>0.5</blocking-threshold>
    <matching-threshold>1</matching-threshold>
    <weight>8</weight>
    <algorithm>org.emau.icmvc.ttp.deduplication.impl.LevenshteinAlgorithm</algorithm>
    <multiple-values>
      <separator> </separator>
      <penalty-not-a-perfect-match>0.5</penalty-not-a-perfect-match>
      <penalty-one-short>0.5</penalty-one-short>
      <penalty-both-short>0.5</penalty-both-short>
    </multiple-values>
  </field>
  <field>
    <name>lastName</name>
    <matching-threshold>0.6</matching-threshold>
    <weight>5</weight>
    <algorithm>org.emau.icmvc.ttp.deduplication.impl.LevenshteinAlgorithm</algorithm>
  </field>
  <field>
    <name>gender</name>
    <matching-threshold>0.75</matching-threshold>
    <weight>3</weight>
    <algorithm>org.emau.icmvc.ttp.deduplication.impl.LevenshteinAlgorithm</algorithm>
  </field>
  <field>
    <name>birthDate</name>
    <blocking-threshold>0.7</blocking-threshold>
    <blocking-mode>NUMBERS</blocking-mode>
    <matching-threshold>0.7</matching-threshold>
```

```
37        <weight>12</weight>
38        <algorithm>org.emau.icmvc.ttp.deduplication.impl.
              LevenshteinAlgorithm</algorithm>
39      </field>
40    </matching>
```

Listing B.7: Konfiguration 7

```
<matching>
    <threshold-possible-match>10.5</threshold-possible-match
       >
    <threshold-automatic-match>14.5</threshold-automatic-
       match>
    <use-cemfim>false</use-cemfim>
    <parallel-matching-after>1000</parallel-matching-after>
    <number-of-threads-for-matching>4</number-of-threads-for
       -matching>
    <field>
      <name>firstName</name>
      <blocking-threshold>0.7</blocking-threshold>
      <matching-threshold>0.8</matching-threshold>
      <weight>6</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         ColognePhoneticAlgorithm</algorithm>
      <multiple-values>
        <separator> </separator>
        <penalty-not-a-perfect-match>0.2</penalty-not-a-
           perfect-match>
        <penalty-one-short>0.6</penalty-one-short>
        <penalty-both-short>0.6</penalty-both-short>
      </multiple-values>
    </field>
    <field>
      <name>lastName</name>
      <matching-threshold>0.6</matching-threshold>
      <weight>3</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         ColognePhoneticAlgorithm/algorithm>
    </field>
    <field>
      <name>gender</name>
      <matching-threshold>0.75</matching-threshold>
      <weight>6</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         ColognePhoneticAlgorithm</algorithm>
    </field>
    <field>
      <name>birthDate</name>
      <blocking-threshold>0.9</blocking-threshold>
      <blocking-mode>NUMBERS</blocking-mode>
      <matching-threshold>1</matching-threshold>
```

```
37        <weight>14</weight>
38        <algorithm>org.emau.icmvc.ttp.deduplication.impl.
              CologneAlgorithm</algorithm>
39      </field>
40    </matching>
```

Listing B.8: Konfiguration 8

```
<matching>
    <threshold-possible-match>12.99</threshold-possible-
       match>
    <threshold-automatic-match>14.5</threshold-automatic-
       match>
    <use-cemfim>false</use-cemfim>
    <parallel-matching-after>1000</parallel-matching-after>
    <number-of-threads-for-matching>4</number-of-threads-for
       -matching>
    <field>
      <name>firstName</name>
      <blocking-threshold>1</blocking-threshold>
      <matching-threshold>1</matching-threshold>
      <weight>8</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         DeterministicAlgorithm</algorithm>
      <multiple-values>
        <separator> </separator>
        <penalty-not-a-perfect-match>0.5</penalty-not-a-
           perfect-match>
        <penalty-one-short>0.5</penalty-one-short>
        <penalty-both-short>0.5</penalty-both-short>
      </multiple-values>
    </field>
    <field>
      <name>lastName</name>
      <matching-threshold>1</matching-threshold>
      <weight>6</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         DeterministicAlgorithm</algorithm>
    </field>
    <field>
      <name>gender</name>
      <matching-threshold>1</matching-threshold>
      <weight>12</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         DeterministicAlgorithm</algorithm>
    </field>
    <field>
      <name>birthDate</name>
      <blocking-threshold>1</blocking-threshold>
      <blocking-mode>NUMBERS</blocking-mode>
      <matching-threshold>1</matching-threshold>
```

```
37        <weight>14</weight>
38        <algorithm>org.emau.icmvc.ttp.deduplication.impl.
              DeterministicAlgorithm</algorithm>
39      </field>
40    </matching>
```

Listing B.9: Konfiguration 9

```
<matching>
    <threshold-possible-match>10.5</threshold-possible-match
       >
    <threshold-automatic-match>14.5</threshold-automatic-
       match>
    <use-cemfim>false</use-cemfim>
    <parallel-matching-after>1000</parallel-matching-after>
    <number-of-threads-for-matching>4</number-of-threads-for
       -matching>
    <field>
      <name>firstName</name>
      <blocking-threshold>0.5</blocking-threshold>
      <matching-threshold>1</matching-threshold>
      <weight>8</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         LevenshteinAlgorithm</algorithm>
      <multiple-values>
        <separator> </separator>
        <penalty-not-a-perfect-match>0.5</penalty-not-a-
           perfect-match>
        <penalty-one-short>0.5</penalty-one-short>
        <penalty-both-short>0.5</penalty-both-short>
      </multiple-values>
    </field>
    <field>
      <name>lastName</name>
      <matching-threshold>0.6</matching-threshold>
      <weight>5</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         LevenshteinAlgorithm</algorithm>
    </field>
    <field>
      <name>gender</name>
      <matching-threshold>0.75</matching-threshold>
      <weight>3</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         LevenshteinAlgorithm</algorithm>
    </field>
    <field>
      <name>birthDate</name>
      <blocking-threshold>0.7</blocking-threshold>
      <blocking-mode>NUMBERS</blocking-mode>
      <matching-threshold>0.7</matching-threshold>
```

```
37        <weight>12</weight>
38        <algorithm>org.emau.icmvc.ttp.deduplication.impl.
              LevenshteinAlgorithm</algorithm>
39      </field>
40      <field>
41        <name>value3</name>
42        <matching-threshold>0.75</matching-threshold>
43        <weight>3</weight>
44        <algorithm>org.emau.icmvc.ttp.deduplication.impl.
              LevenshteinAlgorithm</algorithm>
45      </field>
46    </matching>
```

Listing B.10: Konfiguration 10

```
<matching>
    <threshold-possible-match>13.99</threshold-possible-
       match>
    <threshold-automatic-match>17.5</threshold-automatic-
       match>
    <use-cemfim>false</use-cemfim>
    <parallel-matching-after>1000</parallel-matching-after>
    <number-of-threads-for-matching>4</number-of-threads-for
       -matching>
    <field>
      <name>firstName</name>
      <blocking-threshold>0.7</blocking-threshold>
      <matching-threshold>0.9</matching-threshold>
      <weight>9</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         ColognePhoneticAlgorithm</algorithm>
      <multiple-values>
        <separator> </separator>
        <penalty-not-a-perfect-match>0.2</penalty-not-a-
           perfect-match>
        <penalty-one-short>0.6</penalty-one-short>
        <penalty-both-short>0.6</penalty-both-short>
      </multiple-values>
    </field>
    <field>
      <name>lastName</name>
      <matching-threshold>0.5</matching-threshold>
      <weight>4</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         ColognePhoneticAlgorithm</algorithm>
    </field>
    <field>
      <name>gender</name>
      <matching-threshold>0.75</matching-threshold>
      <weight>6</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         ColognePhoneticAlgorithm</algorithm>
    </field>
    <field>
      <name>birthDate</name>
      <blocking-threshold>0.9</blocking-threshold>
      <blocking-mode>NUMBERS</blocking-mode>
      <matching-threshold>1</matching-threshold>
```

```
37        <weight>12</weight>
38        <algorithm>org.emau.icmvc.ttp.deduplication.impl.
              CologneAlgorithm</algorithm>
39      </field>
40      <field>
41        <name>value3</name>
42        <matching-threshold>0.6</matching-threshold>
43        <weight>9</weight>
44        <algorithm>org.emau.icmvc.ttp.deduplication.impl.
              CologneAlgorithm</algorithm>
45      </field>
46    </matching>
```

Listing B.11: Konfiguration 11

```
<matching>
    <threshold-possible-match>12.99</threshold-possible-
       match>
    <threshold-automatic-match>14.5</threshold-automatic-
       match>
    <use-cemfim>false</use-cemfim>
    <parallel-matching-after>1000</parallel-matching-after>
    <number-of-threads-for-matching>4</number-of-threads-for
       -matching>
    <field>
      <name>firstName</name>
      <blocking-threshold>1</blocking-threshold>
      <matching-threshold>1</matching-threshold>
      <weight>8</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         DeterministicAlgorithm</algorithm>
      <multiple-values>
        <separator> </separator>
        <penalty-not-a-perfect-match>0.5</penalty-not-a-
           perfect-match>
        <penalty-one-short>0.5</penalty-one-short>
        <penalty-both-short>0.5</penalty-both-short>
      </multiple-values>
    </field>
    <field>
      <name>lastName</name>
      <matching-threshold>1</matching-threshold>
      <weight>6</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         DeterministicAlgorithm</algorithm>
    </field>
    <field>
      <name>gender</name>
      <matching-threshold>1</matching-threshold>
      <weight>12</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         DeterministicAlgorithm</algorithm>
    </field>
    <field>
      <name>birthDate</name>
      <blocking-threshold>1</blocking-threshold>
      <blocking-mode>NUMBERS</blocking-mode>
      <matching-threshold>1</matching-threshold>
```

```
37        <weight>14</weight>
38        <algorithm>org.emau.icmvc.ttp.deduplication.impl.
              DeterministicAlgorithm</algorithm>
39      </field>
40      <field>
41        <name>value3</name>
42        <matching-threshold>1</matching-threshold>
43        <weight>9</weight>
44        <algorithm>org.emau.icmvc.ttp.deduplication.impl.
              DeterministicAlgorithm</algorithm>
45      </field>
46    </matching>
```

Listing B.12: Konfiguration 12

```
<matching>
    <threshold-possible-match>2.99</threshold-possible-match
       >
    <threshold-automatic-match>14.5</threshold-automatic-
       match>
    <use-cemfim>false</use-cemfim>
    <parallel-matching-after>1000</parallel-matching-after>
    <number-of-threads-for-matching>4</number-of-threads-for
       -matching>
    <field>
      <name>firstName</name>
      <blocking-threshold>0.4</blocking-threshold>
      <matching-threshold>0.8</matching-threshold>
      <weight>8</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         DeterministicAlgorithm</algorithm>
      <multiple-values>
        <separator> </separator>
        <penalty-not-a-perfect-match>0.1</penalty-not-a-
           perfect-match>
        <penalty-one-short>0.2</penalty-one-short>
        <penalty-both-short>0.2</penalty-both-short>
      </multiple-values>
    </field>
    <field>
      <name>lastName</name>
      <matching-threshold>0.8</matching-threshold>
      <weight>6</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         DeterministicAlgorithm</algorithm>
    </field>
    <field>
      <name>gender</name>
      <matching-threshold>0.75</matching-threshold>
      <weight>3</weight>
      <algorithm>org.emau.icmvc.ttp.deduplication.impl.
         DeterministicAlgorithm</algorithm>
    </field>
    <field>
      <name>birthDate</name>
      <blocking-threshold>0.6</blocking-threshold>
      <blocking-mode>NUMBERS</blocking-mode>
      <matching-threshold>1</matching-threshold>
```

```
37        <weight>9</weight>
38        <algorithm>org.emau.icmvc.ttp.deduplication.impl.
              DeterministicAlgorithm</algorithm>
39      </field>
40    </matching>
```

Listing B.13: Standardkonfiguration

C. Konfigurationen von FRIL

```
<?xml version="1.0" encoding="UTF-8"?>
<configuration deduplication="true">
    <left-data-source class="cdc.impl.datasource.text.
       CSVDataSource" name="sourceA">
        <params>
            <param name="column-separator" value=";"/>
            <param name="source-name" value="sourceA"/>
            <param name="input-file" value="C:\Users\ailee\
               OneDrive\Dokumente\test.csv"/>
        </params>
        <row-model>
            <column column="firstName"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="firstName">
                <empty-values/>
                <params/>
            </column>
            <column column="lastName"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="lastName">
                <empty-values/>
                <params/>
            </column>
            <column column="gender"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="gender">
                <empty-values/>
                <params/>
            </column>
            <column column="birthDate"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="birthDate">
                <empty-values/>
                <params/>
            </column>
```

```
30          </row-model>
31          <preprocessing>
32              <deduplication>
33                  <deduplication-condition acceptance-level="
                        97">
34                      <condition class="cdc.impl.distance.
                            EditDistance"
35                          column="firstName" weight="30">
36                          <params>
37                              <param name="math-level-end"
                                    value="0.5"/>
38                              <param name="match-level-start"
                                    value="0.5"/>
39                          </params>
40                      </condition>
41                      <condition class="cdc.impl.distance.
                            EditDistance"
42                          column="lastName" weight="20">
43                          <params>
44                              <param name="math-level-end"
                                    value="0.4"/>
45                              <param name="match-level-start"
                                    value="0.2"/>
46                          </params>
47                      </condition>
48                      <condition class="cdc.impl.distance.
                            EditDistance"
49                          column="gender" weight="20">
50                          <params>
51                              <param name="math-level-end"
                                    value="0.5"/>
52                              <param name="match-level-start"
                                    value="0.5"/>
53                          </params>
54                      </condition>
55                      <condition class="cdc.impl.distance.
                            EditDistance"
56                          column="birthDate" weight="30">
57                          <params>
58                              <param name="math-level-end"
                                    value="0.4"/>
59                              <param name="match-level-start"
                                    value="0.3"/>
```

```
60                          </params>
61                      </condition>
62                  </deduplication-condition>
63                  <hashing-function columns="birthDate,
                        birthDate" hash="equality"/>
64                  <dedupe-file file="deduplicated-source.csv"/
                        >
65              </deduplication>
66          </preprocessing>
67      </left-data-source>
68  </configuration>
```

Listing C.1: Konfiguration 1

```
<?xml version="1.0" encoding="UTF-8"?>
<configuration deduplication="true">
    <left-data-source class="cdc.impl.datasource.text.
       CSVDataSource" name="sourceA">
        <params>
            <param name="column-separator" value=";"/>
            <param name="source-name" value="sourceA"/>
            <param name="input-file" value="C:\Users\ailee\
               OneDrive\Dokumente\test.csv"/>
        </params>
        <row-model>
            <column column="firstName"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="firstName">
                <empty-values/>
                <params/>
            </column>
            <column column="lastName"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="lastName">
                <empty-values/>
                <params/>
            </column>
            <column column="gender"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="gender">
                <empty-values/>
                <params/>
            </column>
            <column column="birthDate"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="birthDate">
                <empty-values/>
                <params/>
            </column>
        </row-model>
        <preprocessing>
            <deduplication>
                <deduplication-condition acceptance-level="
                   98">
                    <condition class="cdc.impl.distance.
                       SoundexDistance"
                        column="firstName" weight="30">
```

```
36                      <params>
37                          <param name="match-level-start"
                               value="0.5"/>
38                          <param name="math-level-end"
                               value="0.5"/>
39                          <param name="soundex-length"
                               value="5"/>
40                      </params>
41                  </condition>
42                  <condition class="cdc.impl.distance.
                       SoundexDistance"
43                      column="lastName" weight="20">
44                      <params>
45                          <param name="match-level-start"
                               value="0.2"/>
46                          <param name="math-level-end"
                               value="0.4"/>
47                          <param name="soundex-length"
                               value="5"/>
48                      </params>
49                  </condition>
50                  <condition class="cdc.impl.distance.
                       SoundexDistance"
51                      column="gender" weight="20">
52                      <params>
53                          <param name="match-level-start"
                               value="0.5"/>
54                          <param name="math-level-end"
                               value="0.5"/>
55                          <param name="soundex-length"
                               value="5"/>
56                      </params>
57                  </condition>
58                  <condition class="cdc.impl.distance.
                       SoundexDistance"
59                      column="birthDate" weight="30">
60                      <params>
61                          <param name="match-level-start"
                               value="0.3"/>
62                          <param name="math-level-end"
                               value="0.4"/>
63                          <param name="soundex-length"
                               value="5"/>
```

```
64                          </params>
65                      </condition>
66                  </deduplication-condition>
67                  <hashing-function columns="birthDate,
                        birthDate" hash="equality"/>
68                  <dedupe-file file="deduplicated-source.csv"/
                        >
69              </deduplication>
70          </preprocessing>
71      </left-data-source>
72  </configuration>
```

Listing C.2: Konfiguration 2

```
<?xml version="1.0" encoding="UTF-8"?>
<configuration deduplication="true">
    <left-data-source class="cdc.impl.datasource.text.
       CSVDataSource" name="sourceA">
        <params>
            <param name="column-separator" value=";"/>
            <param name="source-name" value="sourceA"/>
            <param name="input-file" value="C:\Users\ailee\
               OneDrive\Dokumente\test.csv"/>
        </params>
        <row-model>
            <column column="firstName"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="firstName">
                <empty-values/>
                <params/>
            </column>
            <column column="lastName"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="lastName">
                <empty-values/>
                <params/>
            </column>
            <column column="gender"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="gender">
                <empty-values/>
                <params/>
            </column>
            <column column="birthDate"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="birthDate">
                <empty-values/>
                <params/>
            </column>
        </row-model>
        <preprocessing>
            <deduplication>
                <deduplication-condition acceptance-level="
                   100">
                    <condition
                        class="cdc.impl.distance.
                           EqualFieldsDistance"
```

```
36                          column="firstName" weight="30">
37                          <params/>
38                      </condition>
39                      <condition
40                          class="cdc.impl.distance.
                               EqualFieldsDistance"
41                          column="lastName" weight="20">
42                          <params/>
43                      </condition>
44                      <condition
45                          class="cdc.impl.distance.
                               EqualFieldsDistance"
46                          column="gender" weight="20">
47                          <params/>
48                      </condition>
49                      <condition
50                          class="cdc.impl.distance.
                               EqualFieldsDistance"
51                          column="birthDate" weight="30">
52                          <params/>
53                      </condition>
54                  </deduplication-condition>
55                  <hashing-function columns="birthDate,
                        birthDate" hash="equality"/>
56                  <dedupe-file file="deduplicated-source.csv"/
                        >
57              </deduplication>
58          </preprocessing>
59      </left-data-source>
60  </configuration>
```

Listing C.3: Konfiguration 3

```
<?xml version="1.0" encoding="UTF-8"?>
<configuration deduplication="true">
    <left-data-source class="cdc.impl.datasource.text.
       CSVDataSource" name="sourceA">
        <params>
            <param name="column-separator" value=";"/>
            <param name="source-name" value="sourceA"/>
            <param name="input-file" value="C:\Users\ailee\
               OneDrive\Dokumente\test.csv"/>
        </params>
        <row-model>
            <column column="firstName"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="firstName">
                <empty-values/>
                <params/>
            </column>
            <column column="lastName"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="lastName">
                <empty-values/>
                <params/>
            </column>
            <column column="gender"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="gender">
                <empty-values/>
                <params/>
            </column>
            <column column="birthDate"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="birthDate">
                <empty-values/>
                <params/>
            </column>
            <column column="plz"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="plz">
                <empty-values/>
                <params/>
            </column>
        </row-model>
        <preprocessing>
```

```
37              <deduplication>
38                  <deduplication-condition acceptance-level="
                        98">
39                      <condition class="cdc.impl.distance.
                            EditDistance"
40                          column="firstName" weight="25">
41                          <params>
42                              <param name="math-level-end"
                                    value="0.5"/>
43                              <param name="match-level-start"
                                    value="0.5"/>
44                          </params>
45                      </condition>
46                      <condition class="cdc.impl.distance.
                            EditDistance"
47                          column="lastName" weight="20">
48                          <params>
49                              <param name="math-level-end"
                                    value="0.4"/>
50                              <param name="match-level-start"
                                    value="0.2"/>
51                          </params>
52                      </condition>
53                      <condition class="cdc.impl.distance.
                            EditDistance"
54                          column="gender" weight="18">
55                          <params>
56                              <param name="math-level-end"
                                    value="0.1"/>
57                              <param name="match-level-start"
                                    value="0.1"/>
58                          </params>
59                      </condition>
60                      <condition class="cdc.impl.distance.
                            EditDistance"
61                          column="birthDate" weight="30">
62                          <params>
63                              <param name="math-level-end"
                                    value="0.4"/>
64                              <param name="match-level-start"
                                    value="0.3"/>
65                          </params>
66                      </condition>
```

```
67                      <condition class="cdc.impl.distance.
                           EditDistance"
68                           column="plz" weight="7">
69                           <params>
70                               <param name="math-level-end"
                                    value="0.1"/>
71                               <param name="match-level-start"
                                    value="0.1"/>
72                           </params>
73                      </condition>
74                  </deduplication-condition>
75                  <hashing-function columns="birthDate,
                        birthDate" hash="equality"/>
76                  <dedupe-file file="deduplicated-source.csv"/
                        >
77              </deduplication>
78          </preprocessing>
79      </left-data-source>
80  </configuration>
```

Listing C.4: Konfiguration 4

```
<?xml version="1.0" encoding="UTF-8"?>
<configuration deduplication="true">
    <left-data-source class="cdc.impl.datasource.text.
       CSVDataSource" name="sourceA">
        <params>
            <param name="column-separator" value=";"/>
            <param name="source-name" value="sourceA"/>
            <param name="input-file" value="C:\Users\ailee\
               OneDrive\Dokumente\test.csv"/>
        </params>
        <row-model>
            <column column="firstName"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="firstName">
                <empty-values/>
                <params/>
            </column>
            <column column="lastName"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="lastName">
                <empty-values/>
                <params/>
            </column>
            <column column="gender"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="gender">
                <empty-values/>
                <params/>
            </column>
            <column column="birthDate"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="birthDate">
                <empty-values/>
                <params/>
            </column>
            <column column="plz"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="plz">
                <empty-values/>
                <params/>
            </column>
        </row-model>
        <preprocessing>
```

```
37              <deduplication>
38                  <deduplication-condition acceptance-level="
                        98">
39                      <condition class="cdc.impl.distance.
                            SoundexDistance"
40                          column="firstName" weight="25">
41                          <params>
42                              <param name="match-level-start"
                                    value="0.5"/>
43                              <param name="math-level-end"
                                    value="0.5"/>
44                              <param name="soundex-length"
                                    value="5"/>
45                          </params>
46                      </condition>
47                      <condition class="cdc.impl.distance.
                            SoundexDistance"
48                          column="lastName" weight="20">
49                          <params>
50                              <param name="match-level-start"
                                    value="0.2"/>
51                              <param name="math-level-end"
                                    value="0.4"/>
52                              <param name="soundex-length"
                                    value="5"/>
53                          </params>
54                      </condition>
55                      <condition class="cdc.impl.distance.
                            SoundexDistance"
56                          column="gender" weight="18">
57                          <params>
58                              <param name="match-level-start"
                                    value="0.1"/>
59                              <param name="math-level-end"
                                    value="0.1"/>
60                              <param name="soundex-length"
                                    value="5"/>
61                          </params>
62                      </condition>
63                      <condition class="cdc.impl.distance.
                            SoundexDistance"
64                          column="birthDate" weight="30">
65                          <params>
```

```
66                                  <param name="match-level-start"
                                        value="0.3"/>
67                                  <param name="math-level-end"
                                        value="0.4"/>
68                                  <param name="soundex-length"
                                        value="5"/>
69                              </params>
70                          </condition>
71                          <condition class="cdc.impl.distance.
                               SoundexDistance"
72                              column="plz" weight="7">
73                              <params>
74                                  <param name="match-level-start"
                                        value="0.1"/>
75                                  <param name="math-level-end"
                                        value="0.1"/>
76                                  <param name="soundex-length"
                                        value="5"/>
77                              </params>
78                          </condition>
79                      </deduplication-condition>
80                      <hashing-function columns="birthDate,
                           birthDate" hash="equality"/>
81                      <dedupe-file file="deduplicated-source.csv"/
                           >
82                  </deduplication>
83              </preprocessing>
84          </left-data-source>
85  </configuration>
```

Listing C.5: Konfiguration 5

```
<?xml version="1.0" encoding="UTF-8"?>
<configuration deduplication="true">
    <left-data-source class="cdc.impl.datasource.text.
       CSVDataSource" name="sourceA">
        <params>
            <param name="column-separator" value=";"/>
            <param name="source-name" value="sourceA"/>
            <param name="input-file" value="C:\Users\ailee\
               OneDrive\Dokumente\test.csv"/>
        </params>
        <row-model>
            <column column="firstName"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="firstName">
                <empty-values/>
                <params/>
            </column>
            <column column="lastName"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="lastName">
                <empty-values/>
                <params/>
            </column>
            <column column="gender"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="gender">
                <empty-values/>
                <params/>
            </column>
            <column column="birthDate"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="birthDate">
                <empty-values/>
                <params/>
            </column>
            <column column="plz"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="plz">
                <empty-values/>
                <params/>
            </column>
        </row-model>
        <preprocessing>
```

```
37              <deduplication>
38                  <deduplication-condition acceptance-level="
                        100">
39                      <condition
40                          class="cdc.impl.distance.
                                EqualFieldsDistance"
41                          column="firstName" weight="25">
42                          <params/>
43                      </condition>
44                      <condition
45                          class="cdc.impl.distance.
                                EqualFieldsDistance"
46                          column="lastName" weight="20">
47                          <params/>
48                      </condition>
49                      <condition
50                          class="cdc.impl.distance.
                                EqualFieldsDistance"
51                          column="gender" weight="18">
52                          <params/>
53                      </condition>
54                      <condition
55                          class="cdc.impl.distance.
                                EqualFieldsDistance"
56                          column="birthDate" weight="30">
57                          <params/>
58                      </condition>
59                      <condition
60                          class="cdc.impl.distance.
                                EqualFieldsDistance"
61                          column="plz" weight="7">
62                          <params/>
63                      </condition>
64                  </deduplication-condition>
65                  <hashing-function columns="birthDate,
                        birthDate" hash="equality"/>
66                  <dedupe-file file="deduplicated-source.csv"/
                        >
67              </deduplication>
68          </preprocessing>
69      </left-data-source>
70  </configuration>
```

Listing C.6: Konfiguration 6

```
<?xml version="1.0" encoding="UTF-8"?>
<configuration deduplication="true">
    <left-data-source class="cdc.impl.datasource.text.
       CSVDataSource" name="sourceA">
        <params>
            <param name="column-separator" value=";"/>
            <param name="source-name" value="sourceA"/>
            <param name="input-file" value="C:\Users\ailee\
               OneDrive\Dokumente\test.csv"/>
        </params>
        <row-model>
            <column column="firstName"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="firstName">
                <empty-values/>
                <params/>
            </column>
            <column column="lastName"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="lastName">
                <empty-values/>
                <params/>
            </column>
            <column column="gender"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="gender">
                <empty-values/>
                <params/>
            </column>
            <column column="birthDate"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="birthDate">
                <empty-values/>
                <params/>
            </column>
        </row-model>
        <preprocessing>
            <deduplication>
                <deduplication-condition acceptance-level="
                   97">
                    <condition class="cdc.impl.distance.
                       EditDistance"
                        column="firstName" weight="30">
```

```
36                          <params>
37                              <param name="math-level-end"
                                    value="0.5"/>
38                              <param name="match-level-start"
                                    value="0.5"/>
39                          </params>
40                      </condition>
41                      <condition class="cdc.impl.distance.
                            EditDistance"
42                          column="lastName" weight="15">
43                          <params>
44                              <param name="math-level-end"
                                    value="0.5"/>
45                              <param name="match-level-start"
                                    value="0.5"/>
46                          </params>
47                      </condition>
48                      <condition class="cdc.impl.distance.
                            EditDistance"
49                          column="gender" weight="25">
50                          <params>
51                              <param name="math-level-end"
                                    value="0.1"/>
52                              <param name="match-level-start"
                                    value="0.1"/>
53                          </params>
54                      </condition>
55                      <condition class="cdc.impl.distance.
                            SoundexDistance"
56                          column="birthDate" weight="30">
57                          <params>
58                              <param name="match-level-start"
                                    value="0.5"/>
59                              <param name="math-level-end"
                                    value="0.5"/>
60                              <param name="soundex-length"
                                    value="5"/>
61                          </params>
62                      </condition>
63                  </deduplication-condition>
64                  <hashing-function columns="birthDate,
                        birthDate" hash="equality"/>
65                  <dedupe-file file="deduplicated-source.csv"/
```

```
                    >
66              </deduplication>
67          </preprocessing>
68      </left-data-source>
69  </configuration>
```

Listing C.7: Konfiguration 7

```
<?xml version="1.0" encoding="UTF-8"?>
<configuration deduplication="true">
    <left-data-source class="cdc.impl.datasource.text.
       CSVDataSource" name="sourceA">
        <params>
            <param name="column-separator" value=";"/>
            <param name="source-name" value="sourceA"/>
            <param name="input-file" value="C:\Users\ailee\
               OneDrive\Dokumente\test.csv"/>
        </params>
        <row-model>
            <column column="firstName"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="firstName">
                <empty-values/>
                <params/>
            </column>
            <column column="lastName"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="lastName">
                <empty-values/>
                <params/>
            </column>
            <column column="gender"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="gender">
                <empty-values/>
                <params/>
            </column>
            <column column="birthDate"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="birthDate">
                <empty-values/>
                <params/>
            </column>
        </row-model>
        <preprocessing>
            <deduplication>
                <deduplication-condition acceptance-level="
                   97">
                 <condition class="cdc.impl.distance.
                    SoundexDistance"
                     column="firstName" weight="30">
```

```
36                      <params>
37                          <param name="match-level-start"
                               value="0.5"/>
38                          <param name="math-level-end"
                               value="0.5"/>
39                          <param name="soundex-length"
                               value="5"/>
40                      </params>
41                  </condition>
42                  <condition class="cdc.impl.distance.
                       SoundexDistance"
43                      column="lastName" weight="15">
44                      <params>
45                          <param name="match-level-start"
                               value="0.5"/>
46                          <param name="math-level-end"
                               value="0.5"/>
47                          <param name="soundex-length"
                               value="5"/>
48                      </params>
49                  </condition>
50                  <condition class="cdc.impl.distance.
                       SoundexDistance"
51                      column="gender" weight="25">
52                      <params>
53                          <param name="match-level-start"
                               value="0.1"/>
54                          <param name="math-level-end"
                               value="0.1"/>
55                          <param name="soundex-length"
                               value="5"/>
56                      </params>
57                  </condition>
58                  <condition class="cdc.impl.distance.
                       EditDistance"
59                      column="birthDate" weight="30">
60                      <params>
61                          <param name="math-level-end"
                               value="0.5"/>
62                          <param name="match-level-start"
                               value="0.5"/>
63                      </params>
64                  </condition>
```

```
65                  </deduplication-condition>
66                  <hashing-function columns="birthDate,
                        birthDate" hash="equality"/>
67                  <dedupe-file file="deduplicated-source.csv"/
                        >
68              </deduplication>
69          </preprocessing>
70      </left-data-source>
71  </configuration>
```

Listing C.8: Konfiguration 8

```
<?xml version="1.0" encoding="UTF-8"?>
<configuration deduplication="true">
    <left-data-source class="cdc.impl.datasource.text.
       CSVDataSource" name="sourceA">
        <params>
            <param name="column-separator" value=";"/>
            <param name="source-name" value="sourceA"/>
            <param name="input-file" value="C:\Users\ailee\
               OneDrive\Dokumente\test.csv"/>
        </params>
        <row-model>
            <column column="firstName"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="firstName">
                <empty-values/>
                <params/>
            </column>
            <column column="lastName"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="lastName">
                <empty-values/>
                <params/>
            </column>
            <column column="gender"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="gender">
                <empty-values/>
                <params/>
            </column>
            <column column="birthDate"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="birthDate">
                <empty-values/>
                <params/>
            </column>
        </row-model>
        <preprocessing>
            <deduplication>
                <deduplication-condition acceptance-level="
                   100">
                 <condition
                     class="cdc.impl.distance.
                        EqualFieldsDistance"
```

```
36                          column="firstName" weight="30">
37                          <params/>
38                      </condition>
39                      <condition
40                          class="cdc.impl.distance.
                               EqualFieldsDistance"
41                          column="lastName" weight="15">
42                          <params/>
43                      </condition>
44                      <condition
45                          class="cdc.impl.distance.
                               EqualFieldsDistance"
46                          column="gender" weight="25">
47                          <params/>
48                      </condition>
49                      <condition
50                          class="cdc.impl.distance.
                               EqualFieldsDistance"
51                          column="birthDate" weight="30">
52                          <params/>
53                      </condition>
54                  </deduplication-condition>
55                  <hashing-function columns="birthDate,
                        birthDate" hash="equality"/>
56                  <dedupe-file file="deduplicated-source.csv"/
                        >
57              </deduplication>
58          </preprocessing>
59      </left-data-source>
60  </configuration>
```

Listing C.9: Konfiguration 9

```
<?xml version="1.0" encoding="UTF-8"?>
<configuration deduplication="true">
    <left-data-source class="cdc.impl.datasource.text.
       CSVDataSource" name="sourceA">
        <params>
            <param name="column-separator" value=";"/>
            <param name="source-name" value="sourceA"/>
            <param name="input-file" value="C:\Users\ailee\
               OneDrive\Dokumente\test.csv"/>
        </params>
        <row-model>
            <column column="firstName"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="firstName">
                <empty-values/>
                <params/>
            </column>
            <column column="lastName"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="lastName">
                <empty-values/>
                <params/>
            </column>
            <column column="gender"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="gender">
                <empty-values/>
                <params/>
            </column>
            <column column="birthDate"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="birthDate">
                <empty-values/>
                <params/>
            </column>
            <column column="plz"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="plz">
                <empty-values/>
                <params/>
            </column>
        </row-model>
        <preprocessing>
```

```
37              <deduplication>
38                  <deduplication-condition acceptance-level="
                        97">
39                      <condition class="cdc.impl.distance.
                            EditDistance"
40                          column="firstName" weight="30">
41                          <params>
42                              <param name="math-level-end"
                                    value="0.4"/>
43                              <param name="match-level-start"
                                    value="0.3"/>
44                          </params>
45                      </condition>
46                      <condition class="cdc.impl.distance.
                            EditDistance"
47                          column="lastName" weight="15">
48                          <params>
49                              <param name="math-level-end"
                                    value="0.5"/>
50                              <param name="match-level-start"
                                    value="0.5"/>
51                          </params>
52                      </condition>
53                      <condition class="cdc.impl.distance.
                            EditDistance"
54                          column="gender" weight="5">
55                          <params>
56                              <param name="math-level-end"
                                    value="0.6"/>
57                              <param name="match-level-start"
                                    value="0.5"/>
58                          </params>
59                      </condition>
60                      <condition class="cdc.impl.distance.
                            EditDistance"
61                          column="birthDate" weight="30">
62                          <params>
63                              <param name="math-level-end"
                                    value="0.6"/>
64                              <param name="match-level-start"
                                    value="0.5"/>
65                          </params>
66                      </condition>
```

```
67                      <condition class="cdc.impl.distance.
                           EditDistance"
68                           column="plz" weight="20">
69                           <params>
70                               <param name="math-level-end"
                                    value="0.9"/>
71                               <param name="match-level-start"
                                    value="0.8"/>
72                           </params>
73                      </condition>
74                  </deduplication-condition>
75                  <hashing-function columns="birthDate,
                        birthDate" hash="equality"/>
76                  <dedupe-file file="deduplicated-source.csv"/
                        >
77              </deduplication>
78          </preprocessing>
79      </left-data-source>
80  </configuration>
```

Listing C.10: Konfiguration 10

```
<?xml version="1.0" encoding="UTF-8"?>
<configuration deduplication="true">
    <left-data-source class="cdc.impl.datasource.text.
       CSVDataSource" name="sourceA">
        <params>
            <param name="column-separator" value=";"/>
            <param name="source-name" value="sourceA"/>
            <param name="input-file" value="C:\Users\ailee\
               OneDrive\Dokumente\test.csv"/>
        </params>
        <row-model>
            <column column="firstName"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="firstName">
                <empty-values/>
                <params/>
            </column>
            <column column="lastName"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="lastName">
                <empty-values/>
                <params/>
            </column>
            <column column="gender"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="gender">
                <empty-values/>
                <params/>
            </column>
            <column column="birthDate"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="birthDate">
                <empty-values/>
                <params/>
            </column>
            <column column="plz"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="plz">
                <empty-values/>
                <params/>
            </column>
        </row-model>
        <preprocessing>
```

```
37              <deduplication>
38                  <deduplication-condition acceptance-level="
                        97">
39                      <condition class="cdc.impl.distance.
                            SoundexDistance"
40                          column="firstName" weight="30">
41                          <params>
42                              <param name="match-level-start"
                                    value="0.3"/>
43                              <param name="math-level-end"
                                    value="0.4"/>
44                              <param name="soundex-length"
                                    value="5"/>
45                          </params>
46                      </condition>
47                      <condition class="cdc.impl.distance.
                            SoundexDistance"
48                          column="lastName" weight="15">
49                          <params>
50                              <param name="match-level-start"
                                    value="0.5"/>
51                              <param name="math-level-end"
                                    value="0.5"/>
52                              <param name="soundex-length"
                                    value="5"/>
53                          </params>
54                      </condition>
55                      <condition class="cdc.impl.distance.
                            SoundexDistance"
56                          column="gender" weight="5">
57                          <params>
58                              <param name="match-level-start"
                                    value="0.5"/>
59                              <param name="math-level-end"
                                    value="0.5"/>
60                              <param name="soundex-length"
                                    value="5"/>
61                          </params>
62                      </condition>
63                      <condition class="cdc.impl.distance.
                            SoundexDistance"
64                          column="birthDate" weight="30">
65                          <params>
```

```
66                                  <param name="match-level-start"
                                        value="0.5"/>
67                                  <param name="math-level-end"
                                        value="0.5"/>
68                                  <param name="soundex-length"
                                        value="5"/>
69                              </params>
70                          </condition>
71                          <condition class="cdc.impl.distance.
                               SoundexDistance"
72                              column="plz" weight="20">
73                              <params>
74                                  <param name="match-level-start"
                                        value="0.8"/>
75                                  <param name="math-level-end"
                                        value="0.9"/>
76                                  <param name="soundex-length"
                                        value="5"/>
77                              </params>
78                          </condition>
79                      </deduplication-condition>
80                      <hashing-function columns="birthDate,
                           birthDate" hash="equality"/>
81                      <dedupe-file file="deduplicated-source.csv"/
                           >
82                  </deduplication>
83              </preprocessing>
84          </left-data-source>
85  </configuration>
```

Listing C.11: Konfiguration 11

```
<?xml version="1.0" encoding="UTF-8"?>
<configuration deduplication="true">
    <left-data-source class="cdc.impl.datasource.text.
       CSVDataSource" name="sourceA">
        <params>
            <param name="column-separator" value=";"/>
            <param name="source-name" value="sourceA"/>
            <param name="input-file" value="C:\Users\ailee\
               OneDrive\Dokumente\test.csv"/>
        </params>
        <row-model>
            <column column="firstName"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="firstName">
                <empty-values/>
                <params/>
            </column>
            <column column="lastName"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="lastName">
                <empty-values/>
                <params/>
            </column>
            <column column="gender"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="gender">
                <empty-values/>
                <params/>
            </column>
            <column column="birthDate"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="birthDate">
                <empty-values/>
                <params/>
            </column>
            <column column="plz"
                converter="cdc.datamodel.converters.
                   DummyConverter" name="plz">
                <empty-values/>
                <params/>
            </column>
        </row-model>
        <preprocessing>
```

```
37              <deduplication>
38                  <deduplication-condition acceptance-level="
                        100">
39                      <condition
40                          class="cdc.impl.distance.
                                EqualFieldsDistance"
41                          column="firstName" weight="30">
42                          <params/>
43                      </condition>
44                      <condition
45                          class="cdc.impl.distance.
                                EqualFieldsDistance"
46                          column="lastName" weight="15">
47                          <params/>
48                      </condition>
49                      <condition
50                          class="cdc.impl.distance.
                                EqualFieldsDistance"
51                          column="gender" weight="5">
52                          <params/>
53                      </condition>
54                      <condition
55                          class="cdc.impl.distance.
                                EqualFieldsDistance"
56                          column="birthDate" weight="30">
57                          <params/>
58                      </condition>
59                      <condition
60                          class="cdc.impl.distance.
                                EqualFieldsDistance"
61                          column="plz" weight="20">
62                          <params/>
63                      </condition>
64                  </deduplication-condition>
65                  <hashing-function columns="birthDate,
                        birthDate" hash="equality"/>
66                  <dedupe-file file="deduplicated-source.csv"/
                        >
67              </deduplication>
68          </preprocessing>
69      </left-data-source>
70  </configuration>
```

Listing C.12: Konfiguration 12