

Aus der Abteilung Versorgungsepidemiologie und Community Health  
(Leiter: Prof. Dr. med. Wolfgang Hoffmann, MPH)  
dem Institut für Community Medicine  
(Direktor: Prof. Dr. med. Wolfgang Hoffmann, MPH)  
der Universitätsmedizin der Ernst-Moritz-Arndt-Universität Greifswald

Thema: Werkzeuggestützte Verfahren für die Realisierung einer Treuhandstelle im  
Rahmen des zentralen Datenmanagements in der epidemiologischen Forschung

Inaugural-Dissertation

zur

Erlangung des akademischen  
Grades

Doktor der Wissenschaften in der Medizin  
(Dr. rer. med.)

der

Universitätsmedizin

der

Ernst-Moritz-Arndt-Universität

Greifswald

2016

vorgelegt von:	Martin Bialke
geb. am:	21. April 1982
in:	Strausberg

Dekan: Prof. Dr. rer. nat. Max P. Baur  
1. Gutachter: Prof. Dr. med. Wolfgang Hoffmann, MPH  
2. Gutachter: Univ.-Prof. Dr. med. Frank Ückert  
Ort, Raum: Institut für Community Medicine ,  
Ellernholzstraße 1/2, 17475 Greifswald  
Hörsaal  
Tag der Disputation: 27. Januar 2017

## Vorbemerkung

Der kumulativen Dissertation liegen folgende Artikel zugrunde:

M. Bialke, T. Bahls, C. Havemann, J. Piegsa, K. Weitmann, T. Wegner und W. Hoffmann, **„MOSAIC. A modular approach to data management in epidemiological studies.“**, *METHODS OF INFORMATION IN MEDICINE*, Bd. 54, Nr. 4, S. 364-371, 8 2015.

M. Bialke, P. Penndorf, T. Wegner, T. Bahls, C. Havemann, J. Piegsa und W. Hoffmann, **„A workflow-driven approach to integrate generic software modules in a Trusted Third Party“**, *Journal of Translational Medicine*, Bd. 13, Nr. 176, 6 2015.

Beide Artikel wurden Open Access veröffentlicht und werden in Übereinstimmung mit den Copyright-Bestimmungen des jeweiligen Verlages im Anhang der Arbeit eingebunden.

---

**Inhalt**

<b>Vorbemerkung</b> .....	<b>I</b>
<b>Inhalt</b> .....	<b>II</b>
<b>Abbildungsverzeichnis</b> .....	<b>IV</b>
<b>Tabellenverzeichnis</b> .....	<b>V</b>
<b>1 Einleitung</b> .....	<b>1</b>
1.1 Hintergrund und Stand der Forschung .....	1
1.2 Fragestellung .....	3
1.3 Gliederung .....	4
<b>2 Methoden</b> .....	<b>5</b>
2.1 Kernkomponenten des zentralen Datenmanagements .....	5
Zentrales Datenmanagement in epidemiologischen Forschungsprojekten.....	5
Beispiele aus der Literatur .....	6
Ableitung wesentlicher Bestandteile .....	9
Unterstützung durch existierende Werkzeuge.....	10
2.2 Aufbau einer Treuhandstelle im Rahmen des zentralen Datenmanagements.....	13
Anforderungen an eine Treuhandstelle aus rechtlicher Sicht .....	13
Rolle der Treuhandstelle im Rahmen eines zentralen Datenmanagements.....	14
Werkzeuggestützter Lösungsansatz .....	15
Übertragbarkeit des Treuhandstellenansatzes.....	16
2.3 Evaluation der Praxistauglichkeit .....	18
Bestimmung des Untersuchungsverfahrens.....	18
Durchführung der Kennzahlenerhebung .....	20
<b>3 Ergebnisse</b> .....	<b>22</b>
3.1 Kernkomponenten des zentralen Datenmanagements .....	22

---

3.2	Werkzeuggestützter Aufbau einer Treuhandstelle im Rahmen des zentralen Datenmanagements .....	23
3.3	Evaluation der Praxistauglichkeit .....	24
<b>4</b>	<b>Diskussion.....</b>	<b>27</b>
4.1	Stärken und Schwächen werkzeuggestützter Verfahren .....	27
4.2	Schlussfolgerungen.....	28
<b>5</b>	<b>Zusammenfassung .....</b>	<b>30</b>
	<b>Literaturverzeichnis .....</b>	<b>31</b>
	<b>Anhang.....</b>	<b>36</b>
<b>A</b>	<b>Wissenschaftliche Artikel .....</b>	<b>36</b>
<b>B</b>	<b>Ergänzungsmaterial.....</b>	<b>52</b>
<b>C</b>	<b>Publikationsliste .....</b>	<b>55</b>
<b>D</b>	<b>Danksagung .....</b>	<b>57</b>

## Abbildungsverzeichnis

<b>Abbildung 1</b> Überblick organisatorischer Aspekte und technischer Maßnahmen im Rahmen eines zentralen Datenmanagements am Beispiel epidemiologischer Kohortenstudien gruppiert nach Studienphasen (Vorbereitungs-, Akquise- und Nutzungsphase) [10].....	6
<b>Abbildung 2</b> Grundlegende Architektur eines zentralen Datenmanagements bestehend aus typischen Komponenten (blau) und Kernkomponenten (orange). Datenschutz und IT-Sicherheit sind unabhängig von den einzelnen Komponenten zu berücksichtigen. (basierend auf [10]) .....	9
<b>Abbildung 3</b> Eine Treuhandstelle (engl. Trusted Third Party, kurz TTP) als Kernelement des zentralen Datenmanagements in der epidemiologischen Forschung [32] .....	14
<b>Abbildung 4</b> Beispiel für eine Treuhandstellenarchitektur (nach [32]) .....	17
<b>Abbildung 5</b> Die Abbildung eines Workflows am Beispiel eines Anwendungsfalls in der Treuhandstelle des Deutschen Zentrums für Herz-Kreislauf-Forschung e.V. (DZHK): personenidentifizierende Daten (IDAT) und die unterzeichnete informierte Einwilligung (IC) des Teilnehmers werden von der Treuhandstelle entgegengenommen, mittels ID-Management E-PIX, Pseudonymisierungswerkzeug gPAS und Einwilligungsmanagement gICS prozessiert und ein Pseudonym (PSN) zur Erfassung der medizinischen Daten (MDAT) an das Studienzentrum zurück übermittelt. [32] .....	24
<b>Abbildung 6</b> Anzahl der mittels E-PIX verwalteten Personen im Projektvergleich (583.079 Personen insgesamt, verwendete Abkürzungen gemäß Tabelle 7). Hinweis: Die Y-Achse wurde für die gewählte Darstellung logarithmisch transformiert. ....	26
<b>Abbildung 7</b> Anzahl der mittels gPAS generierten Pseudonyme im Projektvergleich (2.507.532 Pseudonyme insgesamt, verwendete Abkürzungen gemäß Tabelle 7). Hinweis: Die Y-Achse wurde für die gewählte Darstellung logarithmisch transformiert. ....	26
<b>Abbildung 8</b> Anzahl der mittels gICS verwalteten Einwilligungen im Projektvergleich (68.946 Einwilligungen insgesamt, verwendete Abkürzungen gemäß Tabelle 7). Hinweis: Die Y-Achse wurde für die gewählte Darstellung logarithmisch transformiert. ....	26

## Tabellenverzeichnis

<b>Tabelle 1</b> Anforderungen an ein zentralen Datenmanagement [10] .....	8
<b>Tabelle 2</b> Die Übersicht existierender (Open Source) Lösungen für ein ID-Management listet vorhandene (x) und fehlende (-) Funktionalitäten .....	11
<b>Tabelle 3</b> Die Übersicht existierender (Open Source) Lösungen für Pseudonymisierung und Anonymisierung zeigt vorhandene (x) und fehlende (-) Funktionalitäten .....	11
<b>Tabelle 4</b> Die Übersicht existierender (Open Source) Lösungen für die Verwaltung von Einwilligungen und Widerrufen zeigt vorhandene (x) und fehlende (-) Funktionalitäten.....	12
<b>Tabelle 5</b> Übersicht grundlegender Treuhandstellen-Workflows (nach [32]).....	18
<b>Tabelle 6</b> Erhobene Kennzahlen der Werkzeuge E-PIX, gPAS und gICS.....	21
<b>Tabelle 7</b> Überblick des Einsatzes der MOSAIC-Werkzeuge in der wissenschaftlichen Community [43] (verwendete Abkürzungen: MonDAFIS = Impact of Standardized MONitoring for Detection of Atrial Fibrillation in Ischemic Stroke; DZHK = Deutsches Zentrum für Herz-Kreislauf-Forschung (DZHK) e.V.; TORCH=Translationales Register für Kardiomyopathien; CAU = Christian-Albrechts-Universität zu Kiel; Kifög MV = Summative Evaluation Kindertagesförderungsgesetz Mecklenburg Vorpommern; ZKKR-MV= Zentrales Klinisches Krebsregister MV; GANI_MED = Greifswald Approach to Individualized Medicine; CTMU = vereinbarungsgemäße Struktureinheit der Charité; MVZ = Medizinisches Versorgungszentrum), Stand Januar 2016 .....	25
<b>Tabelle 8</b> Übersicht der von Januar 2015 bis Mai 2016 von den E-PIX-Anwenderprojekten übermittelten Kennzahlen .....	52
<b>Tabelle 9</b> Übersicht der von Januar 2015 bis Mai 2016 von den gPAS-Anwenderprojekten übermittelten Kennzahlen .....	53
<b>Tabelle 10</b> Übersicht der von Januar 2015 bis Mai 2016 von den gICS-Anwenderprojekten übermittelten Kennzahlen. Die Anzahl der Widerrufe wurde erst seit September 2015 erhoben. ....	54

## 1 Einleitung

### 1.1 Hintergrund und Stand der Forschung

Einen wesentlichen Aspekt bei der Durchführung von Kohortenstudien und Registern in der epidemiologischen Forschung stellt das Management von medizinischen Forschungsdaten dar. Nur korrekt erhobene, nachvollziehbar verarbeitete und qualitätsgesicherte Daten können im Rahmen der Datenauswertung für eine valide Beantwortung der wissenschaftlichen Fragestellung herangezogen werden.

Hohe Ansprüche an die Qualität der Daten, besonders bei Einbindung unterschiedlicher Studienzentren, und auch die mit multizentrischen Studien einhergehende standortübergreifende Datenerfassung erfordern ein standardisiertes Vorgehen und die konsequente Gewährleistung des Datenschutzes. Dies empfiehlt den Einsatz eines zentralen Datenmanagements (ZDM).

Grundsätzlich beginnt die Realisierung eines ZDM bereits mit der Schaffung der notwendigen organisatorischen und technischen Voraussetzungen (Klärung von Verantwortlichkeiten, Beachtung von datenschutzrechtlichen und ethischen Erfordernissen, Abstimmung des Forschungsdatensatzes, Bereitstellung von Infrastrukturen, Planung des dauerhaften Betriebs von Infrastrukturen). In der Praxis zeigt sich, dass alle diese Schritte, aber auch die weitere Realisierung des ZDM in Kohortenstudien und Registern, zahlreiche Herausforderungen bergen und stets wiederkehrende organisatorische, prozessbezogene und technische Fragen aufwerfen.

Herausforderungen stellen u.a. die präzise Definition aller Studienvariablen sowie deren Codierung und Beschreibung zum Aufbau eines Data Dictionary dar. Eine webbasierte Datenerfassung setzt die Kenntnis, die Wahl und den Einsatz etablierter Erhebungswerkzeuge als auch die korrekte Realisierung der Erhebungsformulare voraus. Forschungsdaten sollten stets streng getrennt von den Metadaten gespeichert werden. In ausgewählten Szenarien ist zudem die Integration von Gerätedaten in den zentralen Datenbestand erforderlich. Gleichzeitig müssen zahlreiche Anforderungen an die IT-Sicherheit Beachtung finden (Reglementierung des Datenzugriffs, Erstellung von Backups, Datenwiederherstellung im Ernstfall, Risikomanagement) (vgl. [1]). Es ist zu definieren, wie die Qualität der Forschungsdaten kontinuierlich sichergestellt werden kann (Plausibilitätsprüfungen, weitere Qualitätssicherungsmaßnahmen) und welches Zielformat

für den Datenexport bevorzugt wird (SPSS, R, SAS, o.ä.). Falls die Herausgabe der qualitätsgesicherten Forschungsdaten an Dritte beabsichtigt wird, ist zu klären, in welchem Fall dies zulässig ist und wie konkret dabei zu verfahren ist.

Grundsätzlich erfordern Aufbau, Einrichtung, Konfiguration, Betrieb und Wartung der notwendigen Infrastrukturen und Systeme erhebliche zeitliche und finanzielle Aufwände. Kostentreiber sind das erforderliche Personal für die meist notwendigen Anpassungen (Individualisierungen) und der anschließende Betrieb der Systeme. Die Klärung von Zuständigkeiten und Verantwortlichkeiten ist dabei häufig sehr zeitintensiv (Bereitstellung der Infrastrukturen, Verantwortung für Betrieb der Systeme, Kostenübernahme).

Eine wesentliche Herausforderung bei der Umsetzung eines zentralen Datenmanagements stellt die Gewährleistung des Datenschutzes dar. Dies kann in Form einer Treuhandstelle erfolgen [2]. Von Bedeutung ist die Frage nach der Validierung der Datenschutzkonformität gemäß geltender Rechtsprechung (u.a. Bundesdatenschutzgesetz (kurz BDSG), Landesdatenschutzgesetze) in Bezug auf die realisierten technischen, organisatorischen und personellen Maßnahmen. Aus ethischer Sicht ist zu prüfen, wie im Rahmen einer Kohortenstudie bzw. Registers die informierte Einwilligung eines Probanden/Patienten erhoben werden muss. Im Einwilligungsverfahren ist zusätzlich zu definieren, wie im Falle eines Widerrufs im Detail zu verfahren ist. Grundsätzlich gilt: identifizierende Daten (IDAT) sind getrennt von medizinischen Daten (MDAT) zu speichern. MDAT sind stets zum frühestmöglichen Zeitpunkt mithilfe eines geeigneten Werkzeugs zu pseudonymisieren [2].

Darüber hinaus macht der Einsatz mehrerer Studienzentren oft die Zusammenführung von Daten aus unterschiedlichen Datenquellen (u.a. Formulardaten, Laborgerätedaten) und unterschiedlichen Datenformaten mit Hilfe standardisierter ETL<sup>1</sup>-Prozesse erforderlich. Sollen an der Studie teilnehmende Patienten auch über Standortgrenzen hinweg eindeutig identifizierbar bleiben, ist die Verwendung eines zentralen Identitätsmanagements sinnvoll. Dies ist insbesondere dann von Interesse, wenn der Behandlungsverlauf und auch die Nachverfolgbarkeit eines Patienten für die Beantwortung der wissenschaftlichen Fragestellung von Relevanz sind (z.B. im Falle einer Verlegung eines Patienten zwischen teilnehmenden Studienzentren bzw. bei wiederkehrenden Patienten).

---

<sup>1</sup> ETL-Prozess: Prozess zur Extraktion von Daten, deren Anreicherung mit zusätzlichen Informationen, der Transformation der Daten in ein einheitliches Verarbeitungsformat und der abschließenden Überführung der Daten in ein Zielsystem (z.B. Datenbanksystem)

Am Beispiel des deutschen Verbrennungsregisters [3] kann dieser Zusammenhang veranschaulicht werden: Über viele Jahre hinweg wurden Daten von schwerverbrannten Patienten in zahlreichen Registerstandorten mit MS Excel erhoben. Jeder Standort war selbst für die Vergabe einer eindeutigen Datensatzkennung im Zuge der jährlichen Datenzusammenführung verantwortlich. Der Ausschluss von Dopplern war auf diese Weise bisher standortübergreifend nicht möglich und auch die Nachverfolgbarkeit von Patienten war nicht gegeben. Neben der fehlenden zentralen Möglichkeit zur Pseudonymisierung konnten Patienten mangels Einwilligungsverfahren einer Erhebung ihrer Daten nicht widersprechen. Datenschutz- und ethikrelevante Aspekte des Registers wurden nicht in einem Datenschutzkonzept dokumentiert und konnten somit nicht von entsprechender Stelle validiert werden. Diese offenen Punkte schränkten bisher die Untersuchung spezifischer wissenschaftlicher Fragestellungen stark ein [3].

In Summe stellen diese Herausforderungen für zahlreiche Projekte individuelle Nutzungshemmnisse dar. Die Verbesserung der bisher geringen Verbreitung von zentralem Datenmanagement in Kohortenstudien und Registern, das den Verzicht auf damit einhergehende Vorteile in Bezug auf Datenqualität, Datenverfügbarkeit und Datenschutz bedingt, ist daher ein wichtiges transdisziplinäres Forschungsziel.

## **1.2 Fragestellung**

Schwerpunkt dieser Arbeit ist die Untersuchung folgender Fragestellungen:

- I. Was sind Kernkomponenten eines zentralen Datenmanagements und welche Werkzeuge sind für deren Realisierung grundsätzlich erforderlich?
- II. Wie kann der Aufbau einer Treuhandstelle im Rahmen eines zentralen Datenmanagements durch ausgewählte Werkzeuge unterstützt werden?
- III. Aufbauend auf den Ergebnissen der Untersuchung der Fragestellungen I. und II. soll die Tauglichkeit des werkzeuggestützten Ansatzes für die Realisierung eines zentralen Datenmanagements in der epidemiologischen Praxis evaluiert werden.

### 1.3 Gliederung

Die Arbeit ist in drei Abschnitte unterteilt.

- I. Zur Beantwortung der ersten Fragestellung wird zunächst der Begriff „zentrales Datenmanagement“ im Kontext der epidemiologischen Forschung definiert. Ausgehend von beispielhaften epidemiologischen Forschungsprojekten [4, 5] und einem Literaturreview [1, 6, 7] werden typische Anforderungen an ein ZDM ermittelt, um essentielle technische und organisatorische Komponenten im ZDM ableiten zu können. Es wird eine typische Architektur grafisch veranschaulicht. Für ausgewählte Anforderungen erfolgt eine Aufstellung notwendiger Werkzeuge. Funktionalitäten werden mit existierenden Lösungen in der wissenschaftlichen Community abgeglichen und Bedarfslücken identifiziert.
- II. Im zweiten Abschnitt der Arbeit wird auf die Bedeutung der Treuhandstelle im Kontext von ZDM eingegangen und typische „Treuhandstellen“-Komponenten diskutiert. Aus einer Anforderungsanalyse wird abgeleitet, welche Werkzeuge die Etablierung einer Treuhandstelle unterstützen können. Dabei wird besonders der Aspekt der Übertragbarkeit dieser Lösungen für unterschiedliche Studien-/Registerszenarien betrachtet.
- III. Im dritten Abschnitt der Arbeit werden mögliche Methoden zur Untersuchung der Praxistauglichkeit von Werkzeugen für zentrales Datenmanagement in Registern und Kohortenstudien vorgestellt, notwendige Voraussetzungen sowie Vor- und Nachteile untersucht. Im Anschluss wird eine ausgewählte Methode genutzt, um die Praxistauglichkeit der entstandenen Werkzeuge exemplarisch zu prüfen. Zuletzt werden die erzielten Ergebnisse kritisch diskutiert.

## 2 Methoden

### 2.1 Kernkomponenten des zentralen Datenmanagements

Datenmanagement orientiert sich grundsätzlich am Lebenszyklus von Forschungsdaten [8] und umfasst laut [9] „die Erzeugung, Verarbeitung, Speicherung und Nutzbarmachung von Forschungsdaten“. Ein Schwerpunkt des zentralen Datenmanagements im epidemiologischen Kontext liegt auf der datenschutzkonformen Zusammenführung von Forschungsdaten aus heterogenen Datenquellen und häufig aus unterschiedlichen Standorten. Der Begriff zentrales Datenmanagement beschreibt sowohl die dafür erforderlichen organisatorischen und technischen Aufgaben als auch die notwendigen methodischen und konzeptionellen Prozesse. [10]

#### Zentrales Datenmanagement in epidemiologischen Forschungsprojekten

Am *Institut für Community Medicine der Universitätsmedizin Greifswald* wurden und werden zahlreiche epidemiologische Projekte mit zentralem Datenmanagement durchgeführt. Die Realisierung eines zentralen Datenmanagements ist u.a. abhängig von Setting und individuellen, internen Abläufen einer Studie.

Die *SHiP - Studie (Study of Health in Pomerania)* ist eine populationsbasierte Kohortenstudie, deren Ziel es ist, den Gesundheitszustand der Bevölkerung im Nordosten von Deutschland zu charakterisieren. Das Studiendesign und der Untersuchungsumfang der mehr als 14.000 Probanden (1997 – 2012) (vgl. [4, 11]) machen ein effizientes Management, validierbare Konsentierungsverfahren und eine Bioproben-Verwaltung erforderlich. Neben diagnostischen Gerätedaten aus medizinischen Untersuchungen (Blutdruckmessung, EKG, Echokardiographie, Sonographie), Eingaben aus persönlichen computergestützten Interviews (CAPI) und handschriftlichen Fragebögen werden die erhobenen Daten durch eine jährliche Vitalstatus-Abfrage durch die Unabhängige Treuhandstelle am Institut für Community Medicine ergänzt. Die Daten- und Biomaterialübergaben zur Untersuchung wissenschaftlicher Fragestellungen erfolgen nach einem standardisierten Verfahren mit definierten Use und Access-Regeln über eine Transferstelle.

Bei der *GANI\_MED-Forschungsplattform (Greifswald Approach to Individualized Medicine)* steht die heterogene, dezentrale Datenerfassung im Fokus. Es gilt Daten aus unterschiedlichen Quellen, wie klinische Labordaten, MRT-Gerätedaten oder Dentalbefunde, zu erfassen. Zudem stellt die formularbasierte Erhebung auf mobilen und stationären

Endgeräten individuelle Anforderungen an Synchronisationsprozesse. Eine zentrale Verwaltung der digitalen Einwilligungsdokumente (engl. Informed Consent, kurz IC) ermöglicht die Realisierung komplexer und gleichzeitig dynamischer Konsentierungs- und Widerrufsmechanismen (Consent Management). [5]

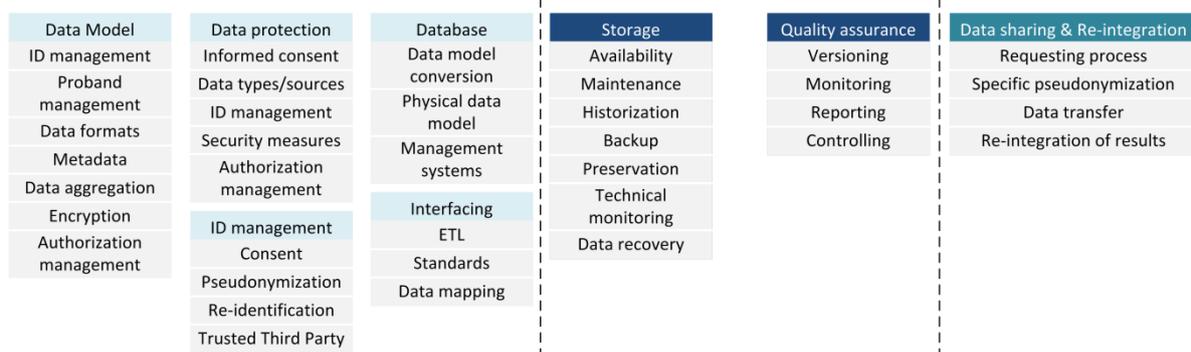
Die *Nationale Kohorte* e.V. [12] ist ein gemeinsames, interdisziplinäres Vorhaben von Wissenschaftlern der Universitäten, der Helmholtz-Gemeinschaft und anderer Forschungsorganisationen in Deutschland. Die Rekrutierung der etwa 200.000 Probanden findet deutschlandweit in 18 Studienzentren in acht geographischen Clustern statt. Eine Besonderheit des zentralen Datenmanagements der Nationalen Kohorte ist, dass sowohl die Datenintegration, als auch die Rollen- und Rechteverwaltung zentral erfolgen. Dabei findet die Verarbeitung der personenidentifizierenden Daten in einer Unabhängigen Treuhandstelle statt (Datenverarbeitung im Auftrag gemäß §11 BDSG).

Abgeleitet aus diesen konkreten Beispielen, werden typische organisatorische Aspekte und technische Maßnahmen eines zentralen Datenmanagements im epidemiologischen Kontext zusammenfassend in Abbildung 1 [10] dargestellt.

#### Organizational issues



#### Technical measures



**Abbildung 1 Überblick organisatorischer Aspekte und technischer Maßnahmen im Rahmen eines zentralen Datenmanagements am Beispiel epidemiologischer Kohortenstudien gruppiert nach Studienphasen (Vorbereitungs-, Akquise- und Nutzungsphase) [10]**

#### Beispiele aus der Literatur

In der Literatur wurden vor allem einzelne Aspekte und Anforderungen an das Datenmanagement in Kohortenstudien/Registern untersucht und diskutiert.

Michaelik et al. [13] erfassen im Projekt *Koreg-IT* grundlegende Anwendungsfälle in Kohortenstudien/Registern. Im Vordergrund der Arbeiten stehen betriebswirtschaftliche und

organisatorische Aspekte von Studienanfang bis Studienende aus Sicht der IT. Es wird bewusst auf konkrete Problemlösungsansätze oder technische Empfehlungen in Bezug auf das Datenmanagement verzichtet, Anforderungen an ein ZDM stehen nicht im Fokus der Betrachtungen.

Sariyar et al. [14] untersuchen Anforderungen in Registern bzgl. Datenqualität, Datenschutz und Datensicherheit. Die Autoren stellen einen Katalog von erforderlichen Maßnahmen und Handlungsanweisungen vor.

Das *Memorandum Register Versorgungsforschung des DNVF* [15] stellt die Beurteilung der Registerqualität in den Vordergrund (u.a. Validierung der Datenerhebung, Datenqualität, Qualitätsanforderungen wie Transparenz, Flexibilität, Anpassungsfähigkeit). Es werden rechtliche Aspekte des Datenschutzes erläutert und Empfehlungen für den Aufbau von Registern genannt. Detaillierte Anforderungen an Datenmanagement werden jedoch nicht erfasst.

Jensen [7] fasst *Leitlinien zum Forschungsdatenmanagement* zusammen. Diese fokussieren den Bereich der Sozialwissenschaften, benennen relevante Themen und präsentieren bereits konkrete Lösungsvorschläge für ein Datenmanagement in Bezug auf die Codierung von Variablen, den Fragebogenentwurf, die Datenqualität, die Datensicherheit, die Versionierung von Daten sowie deren langfristige Archivierung. Darüber hinaus wird auf ausgewählte rechtliche Aspekte des Datenmanagements im Urheberrechtsgesetz (UrhG) und im BDSG, insbesondere zur Anonymisierung und Pseudonymisierung von Daten, eingegangen.

Die *Technologie- und Methodenplattform e.V.* unterstützt die medizinische Verbundforschung beratend, konzeptuell sowie durch zahlreiche Gutachten und ausgewählte Werkzeuge [2, 16, 17, 18, 19, 20, 21]. Pommerening et al. [2] setzen den Fokus epidemiologischer Kohortenstudien und Register auf die langfristige Speicherung qualitätsgesicherter Forschungsdaten und stellen die „*informationelle Gewaltenteilung*“ ([2] S. 48 ff.) als wesentliche Grundlage eines datenschutzkonformen medizinischen Datenmanagements vor. Um den Anforderungen des BDSG Rechnung zu tragen, sind die frühestmögliche Trennung von identifizierenden und medizinischen Daten, die eindeutige Zuordnung von Patienten aus unterschiedlichen Quellen mittels Patientenliste (Record Linkage), eine Pseudonymisierungsmöglichkeit und der Einsatz eines Datentreuhänders zulässige Maßnahmen ([2] S. 37 ff.). Darüber hinaus werden Anforderungen an die informierten Einwilligungen und Widerrufe aufgezeigt ([2] S. 30 ff.)

Meyer et al. [6] fassen Vorteile zentralen Datenmanagements zusammen und ergänzen nicht-funktionale Anforderungen in Bezug auf Erweiterbarkeit, Transparenz, Verfügbarkeit, Integrität, IT-Sicherheit und Datenschutz. Ausgehend von Datenmanagement-Prozessen werden für die Umsetzung eines ZDM geeignete technische Maßnahmen für eine ETL-basierte Datenintegration, ein generisches Datenmodell, webbasierte Erhebungsformulare, eine Plausibilitätsprüfung, ein Rollen- und Rechtemanagement, ein Versionsmanagement, ein geeignetes Monitoring und Methoden zur Datensicherung aufgeführt.

Darüber hinaus sehen Rani et al. [22] die organisierte Datenbereitstellung zur Nachnutzung von bereits qualitätsgesicherten Forschungsdaten als wesentlichen Teil des Datenmanagements an.

Unterscheidung	Anforderung	Beschreibung
<b>Funktional</b>	Erschließung von Datenquellen	Spezifikation von Datenquellen, Datenformaten, Schnittstellen und Übermittlung der Daten.
	Aufbereitung und Integration	Realisierung der ETL-Prozesse, Meta-Datenanreicherung und Sicherung der Datenqualität.
	Standardisierte Speicherung	Konzeption und Realisierung eines generischen Datenmodells und Bereitstellung der notwendigen IT-Infrastruktur.
	Use & Access	Bietet Möglichkeiten zur Datenexploration, Abstimmung des Datenantragsprozesses, als auch den Im- und Export der Daten.
<b>Nicht-Funktional</b>	Integrität	Umfasst Maßnahmen zur Gewährleistung von Datenkonsistenz, -sicherheit und -schutz.
	Transparenz	Überwachung und kontinuierliche Kontrolle von Prozessen und Maßnahmen für vollständige Nachvollziehbarkeit und Reproduzierbarkeit.
	Erweiterbarkeit	Hard- und Software müssen entsprechend Studierweiterungen und wachsenden Anforderungen flexibel erweitert werden können.
	Verfügbarkeit	Erhebung, Speicherung, Archivierung und Bereitstellung der Daten müssen langfristig und ausfallsicher erfolgen.
	Datenschutz	Gemäß den Datenschutzgesetzen des Bundes und der Länder müssen personenidentifizierende und medizinische Daten zum frühestmöglichen Zeitpunkt getrennt werden. Für klardefinierte Use Cases ist eine eindeutige Identifikation von Probanden erforderlich, müssen Einverständnisse, Ermächtigungen und Widerrufe effektiv verwaltet werden und die Pseudonymisierung und Anonymisierung der Daten möglich sein.
	IT- und Informationssicherheit	Umfasst gemäß BSI Grundschriftkatalog [1] Maßnahmen zur Authentifizierung, Autorisierung, gesicherten Datenübertragung, verschlüsselten Datenspeicherung und zur Absicherung der Netzwerkstrukturen gegen unberechtigte Zugriffe.

**Tabelle 1 Anforderungen an ein zentralen Datenmanagement [10]**

Zusammengefasst muss ein ZDM den in Tabelle 1 dargestellten Anforderungen genügen. Dabei sind die funktionalen Anforderungen vor allem durch den Lebenszyklus von Forschungsdaten bedingt [8], aber gleichzeitig auch eng mit der Realisierung der nicht-funktionalen Anforderungen verknüpft. Die Umsetzung der einzelnen Anforderungen erfolgt durch entsprechende organisatorische und technische Maßnahmen. [10]

## Ableitung wesentlicher Bestandteile

Ausgehend von den identifizierten Anforderungen lassen sich essentielle organisatorische und technische Bestandteile eines ZDM ableiten. SHiP, GANI\_MED und die Nationale Kohorte (s.o.) zeigen, dass die Anforderungsausprägung abhängig vom individuellen Szenario variieren kann.

Idealerweise sollten vor allem wiederkehrende Problemstellungen bei der Realisierung eines ZDM durch nachhaltige Lösungsansätze adressiert werden können. Diese Kernkomponenten (vgl. [23, 24]) sind zentrale Bestandteile des ZDM und sollten so modular gestaltet sein, dass sie unabhängig von anderen Bestandteilen verwendet werden können.

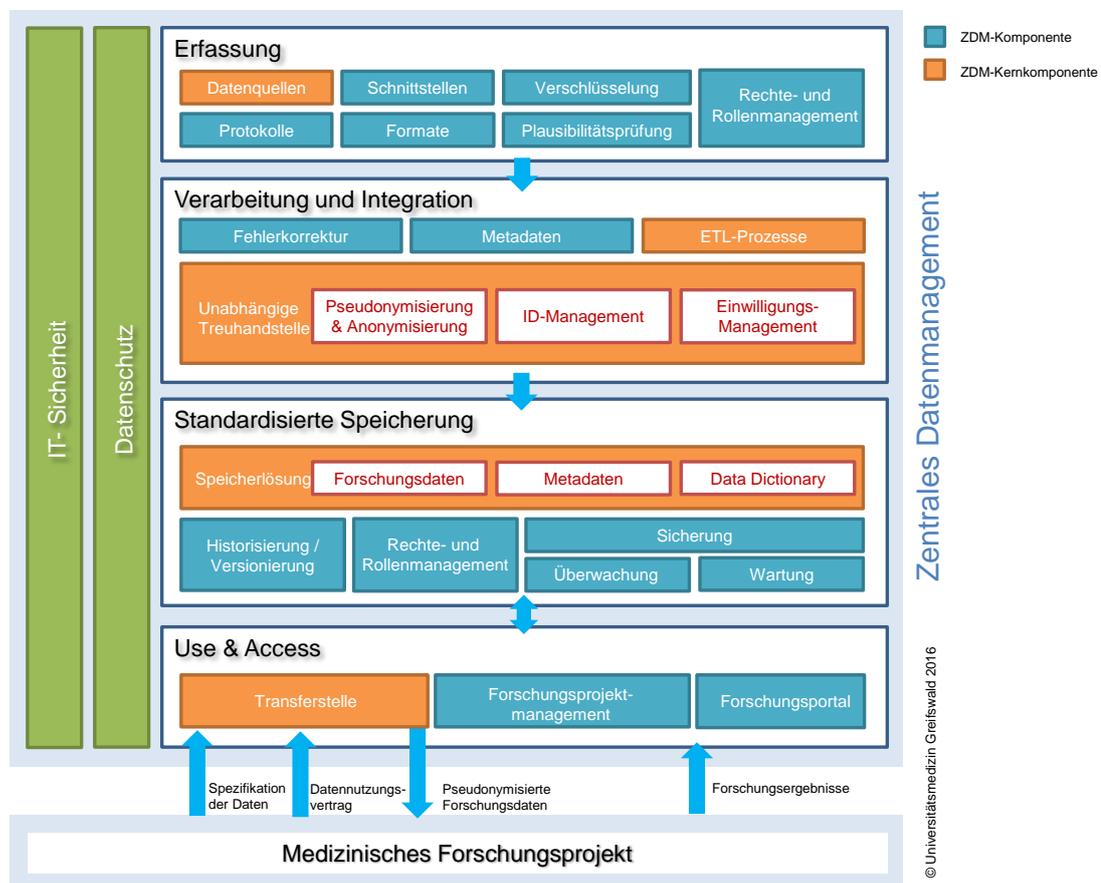


Abbildung 2 Grundlegende Architektur eines zentralen Datenmanagements bestehend aus typischen Komponenten (blau) und Kernkomponenten (orange). Datenschutz und IT-Sicherheit sind unabhängig von den einzelnen Komponenten zu berücksichtigen. (basierend auf [10])

Unabhängig vom konkreten Design und Setting eines epidemiologischen Forschungsprojekts sollte ein ZDM mindestens die Realisierung folgender Kernkomponenten umfassen:

- mindestens eine ausgewählte Datenquelle,
- die Verarbeitung und Anreicherung der erhobenen Daten (ETL-Prozess),

- die Gewährleistung des Datenschutzes durch frühestmögliche Trennung von identifizierenden und medizinischen Daten (Treuhandstelle),
- die standardisierte Speicherlösung für die getrennte Speicherung von pseudonymisierten (medizinischen) Forschungsdaten und zugehörigen Metadaten auf Basis eines abgestimmten Data Dictionary, sowie
- ein einheitliches Use and Access Verfahren zur Bereitstellung, Nutzung und Auswertung der Forschungsdaten (Transferstelle)

In Abbildung 2 werden typische Bestandteile und Kernkomponenten eines ZDM zusammenfassend in Form einer grundlegenden ZDM-Architektur dargestellt.

### **Unterstützung durch existierende Werkzeuge**

Um den Aufwand zur Umsetzung eines ZDM zu minimieren, sollte im Idealfall auf vorhandene (und nach Möglichkeit kostenfreie) Lösungen zurückgegriffen werden. Am Beispiel der Kernkomponente Treuhandstelle (bestehend aus ID-Management, Pseudonymisierung/ Anonymisierung, sowie Einwilligungsmanagement) wurde im Rahmen dieser Arbeit exemplarisch erhoben, welche entsprechenden Werkzeuge in der wissenschaftlichen Community verbreitet sind, die zur Nachnutzung in Frage kommen. Die Ergebnisse dieser Recherche werden nachfolgend überblicksartig dargestellt.

#### *ID-Management*

Um Teilnehmer von Kohortenstudien und Registern auch standortübergreifend eindeutig identifizieren zu können, ist im Rahmen eines ZDM ein einheitliches und konsistentes ID-Management erforderlich. Dieses sollte den Anwender über potentielle Synonymfehler informieren und bei deren Auflösung unterstützen. Vor allem im epidemiologischen Kontext sind Variationen von IDATs (z.B. der Name „Klaus Dieter Müller“ in Studienzentrum A und der Name „Klaus-Diether Müller“ in Studienzentrum B) von entscheidender Bedeutung. Unterschiedliche Schreibweisen von IDATs müssen zwar eindeutig einer real existierenden Person zugeordnet werden können, deren Ausprägung aber in den angeschlossenen Systemen erhalten bleiben, um bspw. Daten aus anderen Quellen jederzeit zuordnen zu können. Um diesem Problem Rechnung zu tragen, muss eine Person mehr als nur eine „korrekte“ Schreibweise von IDAT besitzen können. Hier kann ein Konzept von Haupt- und Nebenidentitäten [25] zur Lösung beitragen. Tabelle 2 zeigt typische funktionale Anforderungen an ein ID-Management und listet entsprechend existierende Werkzeuge auf.

Anforderungen	TMF PID-Generator [18]	Mainzliste [26]	OpenEMPI [27]	E-PIX [28]
Datenzusammenführung aus unterschiedlichen Studienzentren	x	x	x	x
Umgang mit fehlerhaften/unvollständigen Daten	x	x	x	x
Unterstützung für IHE Profile (PIX/PDQ)	-	-	x	Notification fehlt
Verwaltung lokaler Identifier	-	-	x	x
Matching-Algorithmus	Phonetisch	Phonetisch	Diverse	Levenstein-Distanz
Unterstützung beim Auflösen von Synonymfehlern	-	-	x	x
Unterstützung von Nebenidentitäten	-	-	-	x

Tabelle 2 Die Übersicht existierender (Open Source) Lösungen für ein ID-Management listet vorhandene (x) und fehlende (-) Funktionalitäten

### *Pseudonymisierung/Anonymisierung*

Die Verwendung von Pseudonymen (PSN) gestattet, medizinische Daten frei von identifizierenden Daten zu speichern, erlaubt aber gleichzeitig befugtem Personal, den Personenbezug im Bedarfsfall wiederherzustellen. Dies darf dagegen bei anonymisierten medizinischen Daten nicht möglich sein.

Existierende Pseudonymisierungswerkzeuge (vgl. Tabelle 3), wie der *Pseudonymisierungsdienst der TMF<sup>2</sup> (TMF PSD)* [19] und der *generic Pseudonym Administration Service (gPAS)* [29] erzeugen je Eingabewert (z.B. Personenidentifikator) mindestens ein Pseudonym, das zur Speicherung der medizinischen Daten genutzt werden kann. gPAS speichert die Zuordnung von Originalwert und Pseudonym in einer Zuordnungstabelle. Wird der entsprechende Eintrag in der Zuordnungstabelle gelöscht, sind die unter Verwendung des Pseudonyms gespeicherten medizinischen Daten unumkehrbar anonymisiert.

Anforderungen	TMF PSD [19]	gPAS [29]	Anon-Tool [21]
Pseudonymerzeugung	x	x	-
Depseudonymisierung	x	x	-
Pseudonymisierung ohne dauerhafte Speicherung	x	-	-
Beliebige Zeichenketten als Eingabewerte	-	x	-
Nutzung variabler Alphabete und Pseudonymlängen	-	x	-
Prüfzifferalgorithmen	-	x	-
Vergabe von Prä- und Suffixen	-	x	-
Mehrfachpseudonymisierung	-	x	-
Anonymisierung einzelner Datensätze	-	x	x
k-Anonymisierung / I-Diversität	-	-	x

Tabelle 3 Die Übersicht existierender (Open Source) Lösungen für Pseudonymisierung und Anonymisierung zeigt vorhandene (x) und fehlende (-) Funktionalitäten

<sup>2</sup> Technologie- und Methodenplattform für die vernetzte medizinische Forschung e.V.

Das *Anon-Tool* [21] hingegen setzt bei der Anonymisierung auf die geringfügige Änderung der medizinischen Informationen selbst (Generalisierung), denn in Einzelfällen, wie bei seltenen Krankheiten oder unter Zuhilfenahme von Sekundärinformationen, kann es vorkommen, dass das bloße Entfernen der identifizierenden Merkmale einen Datensatz nicht ausreichend unkenntlich macht. Dies kann durch die Transformation ausgewählter Datensatzattribute (k-Anonymisierung<sup>3</sup>, I-Diversität<sup>4</sup>) erreicht werden. Diese Verfahren gehen jedoch mit Informationsverlust einher [21].

### Einwilligungsmanagement

Werkzeuge im Bereich des Einwilligungsmanagements (Consent Management) unterscheiden sich deutlich im Anwendungskontext. Wie in Tabelle 4 dargestellt, unterstützt der *TMF Informed Consent Wizard* [17] vorwiegend bei der Erstellung von Einwilligungsdokumenten. Die *Consent Management Suite (COMS)* [30] unterstützt die Verwaltung von Einwilligungen im Behandlungskontext. Die Stärken des *generic Informed Consent Service (gICS)* [31] werden vor allem im Studienkontext sichtbar. Modulare Einwilligungen und Widerrufe können automatisiert validiert und zudem in unterschiedlichen Studiensettings nachgenutzt werden.

Anforderungen	TMF Informed Consent Wizard [17]	Consent Management Suite COMS [30]	gICS [31]
<b>Hilfestellung bei der Formulierung von Einwilligungen und Widerrufen</b>	x	-	-
Erstellung von Einwilligungen und Widerrufen	x	x	x
<b>Verwaltung von Einwilligungen und Widerrufen</b>	-	x	x
IHE-Unterstützung „Basic Patient Privacy Consent“	-	x	-
<b>Modularisierung von Einwilligungen und Widerrufen für verbesserte Nutzbarkeit und Versionierung</b>	-	-	x
Unterstützung policy-spezifischer, automatisierbarer Abfragen	-	-	x

Tabelle 4 Die Übersicht existierender (Open Source) Lösungen für die Verwaltung von Einwilligungen und Widerrufen zeigt vorhandene (x) und fehlende (-) Funktionalitäten

<sup>3</sup> Die k-Anonymisierung stellt ein Anonymisierungsmodell dar. Je höher k gewählt wird, umso stärker werden Datengruppen (der Größe k) durch Verallgemeinerung von Quasi-Identifikatoren anonymisiert. [48]

<sup>4</sup> Die I-Diversität ist eine Maßzahl der Anonymisierung und optimiert das Prinzip der k-Anonymisierung. Dadurch wird sichergestellt, dass nicht-verallgemeinerte Merkmalsausprägungen einer k-Gruppe in mindestens I Variationen auftreten. [48]

## 2.2 Aufbau einer Treuhandstelle im Rahmen des zentralen Datenmanagements

### Anforderungen an eine Treuhandstelle aus rechtlicher Sicht

Beim Aufbau eines ZDM für Kohortenstudien und Register erfordert die Erhebung, Verarbeitung und Speicherung personenbezogener medizinischer Daten die Einhaltung weitreichender nationaler und internationaler datenschutzrechtlicher Rahmenbedingungen [32]. Dazu zählen in Deutschland die Landesdatenschutzgesetze, das BDSG [33], das Übereinkommen zum Schutz des Menschen bei der automatischen Verarbeitung personenbezogener Daten [34] sowie im internationalen Kontext ethische Grundsätze für die medizinische Forschung am Menschen (Deklaration von Helsinki) [35].

Daraus resultierende Anforderungen wurden von der TMF bereits in der entsprechenden Leitlinie [2] überblicksartig erarbeitet. Demnach sind typische datenschutzrechtliche und ethische Bedingungen [32] u.a. eine in der Regel vom Teilnehmer selbst unterzeichnete Einwilligungserklärung als Grundlage der Forschungsdatenerhebung, die Datenvermeidung und -sparsamkeit gemäß §3a BDSG im gesamten Erhebungs-, Speicherungs- und Auswertungsprozess [33] sowie die frühestmögliche Trennung und getrennte Speicherung von medizinischen und personenidentifizierenden Daten gemäß §40 BDSG.

Der TMF Leitfaden [2] weist darauf hin, dass wesentliche Herausforderungen des Datenschutzes durch Etablierung einer Treuhandstelle bewältigt werden können. Laut [2] stellt eine Treuhandstelle eine „Zusammenstellung technischer und organisatorischer Maßnahmen zur Gewährleistung grundlegender Anforderungen an Datenschutz und IT-Sicherheit“ [32] dar. Dabei sind folgende grundsätzlichen Anforderungen zu berücksichtigen [2]:

- eine informationelle Gewaltenteilung aus technischer und organisatorischer Sicht (Trennung von IDAT und MDAT),
- ein elektronisches ID-Management,
- ein sicherer Pseudonymisierungsalgorithmus,
- eine Wahrung der Rechte des Teilnehmers durch Einsatz von differenzierten Einwilligungen und Widerrufern (u.a. zur Gewährleistung von Recht auf Wissen, Recht auf Nichtwissen, Recht auf Widerspruch), und
- eine rechtliche, personelle und räumliche Eigenständigkeit und Unabhängigkeit (garantiert durch vertragliche Vereinbarungen (§28 BDSG) und die Übertragung sämtlicher Verantwortung zur Datenverarbeitung (Funktionsübertragung))

Die Aufwände zur Realisierung einer Treuhandstelle im Rahmen eines zentralen Datenmanagements sind abhängig vom Setting der Studie bzw. des Registers [32]. Im Einzelfall sind zahlreiche ergänzende Maßnahmen erforderlich, wie u.a. die Reglementierung des Zugriffs, die Umsetzung von Zugangsbeschränkungen sowie die Einrichtung getrennter Netzwerkzonen.

### Rolle der Treuhandstelle im Rahmen eines zentralen Datenmanagements

Wesentlicher Bestandteil bei der Implementierung eines ZDM für die epidemiologische Forschung ist die Realisierung einer unabhängigen Treuhandstelle [32]. Laut Havemann et al. [36] umfassen zentrale Verantwortlichkeiten einer Treuhandstelle:

- das ID-Management: die Verwaltung und Zuordnung von IDAT der Studienteilnehmer (Dublettenprüfung und –bereinigung) aus unterschiedlichen beteiligten Systemen (u.a. Studienzentren, Krankenhausinformationssysteme),
- die Verwaltung von Einwilligungen und Widerrufen,
- die Pseudonymisierung und Depseudonymisierung von IDAT, sowie
- die Unterstützung zahlreicher studienrelevanter Prozesse, wie die Patienten-Kontaktierung, die Datenherausgabe oder die Durchführung von Follow-Ups [36]

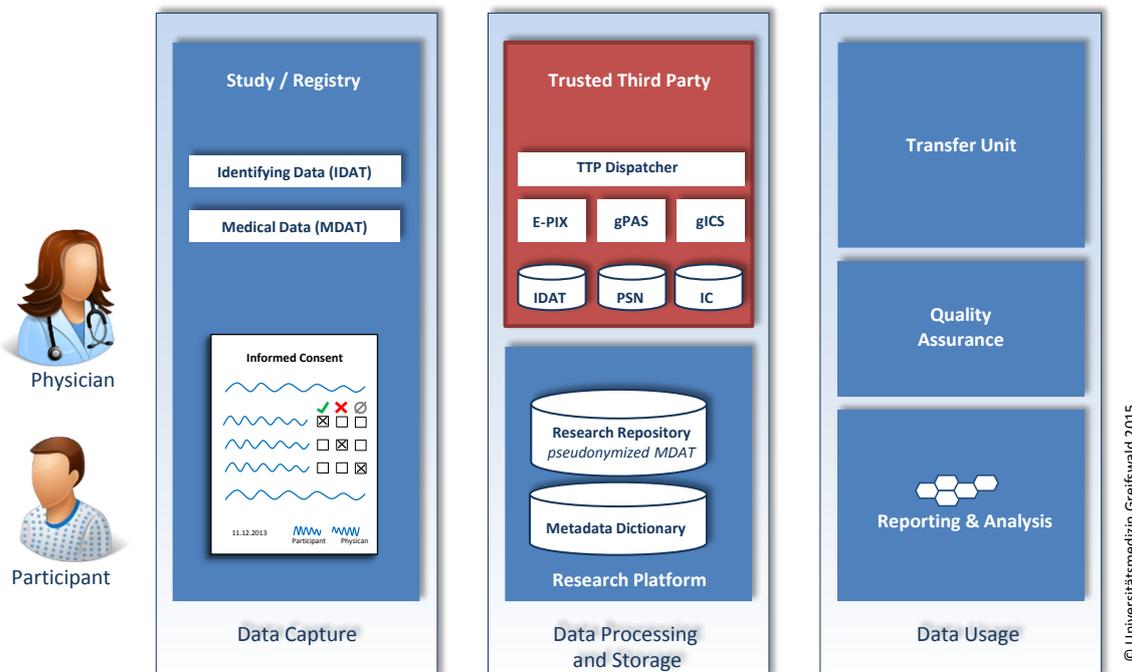


Abbildung 3 Eine Treuhandstelle (engl. Trusted Third Party, kurz TTP) als Kernelement des zentralen Datenmanagements in der epidemiologischen Forschung [32]

Dabei werden identifizierende Daten ausschließlich in einer Treuhandstelle gespeichert, während entsprechend pseudonymisierte medizinische Daten in einer separaten Forschungsdatenbank (außerhalb der Treuhandstelle) abgelegt werden. Die Nutzung der pseudonymisierten Daten erfolgt im Rahmen der Datenauswertung, der Berichterstellung, der Datenbereitstellung zur Sekundärnutzung, sowie der Sicherung der Datenqualität. Dieser Zusammenhang wird in Abbildung 3 [32] veranschaulicht.

### **Werkzeuggestützter Lösungsansatz**

Der Aufbau einer Treuhandstelle im Rahmen eines ZDM kann, ausgerichtet am Datenschutz-Leitfaden der TMF [2], mithilfe der frei verfügbaren Werkzeuge *E-PIX* (*Enterprise Patient Identifier Cross Referencing*), *gPAS* und *gICS* (vgl. Tabelle 2 – 4) bewerkstelligt werden [32]. Diese modularen Werkzeuge bieten den erforderlichen Funktionsumfang, folgen grundsätzlich dem Prinzip „*Privacy by Design*“ [37] und setzen auf einheitliche technische Standards, wie beispielsweise eine Java-basierte und serviceorientierte Architektur. Gleichzeitig unterstützen diese die informationelle Gewaltenteilung und adressieren zentrale Verantwortlichkeiten eines Datentreuhänders.

Der *E-PIX* stellt sicher, dass jeder Studienteilnehmer auch im Fall von unvollständigen oder fehlerhaften demografischer Informationen eindeutig identifiziert werden kann. Wurde ein Patient bspw. in einem angeschlossenen System potentiell doppelt angelegt, unterstützt der *E-PIX* den Datentreuhänder grafisch bei der Identifikation und Bereinigung dieses möglichen Synonymfehlers.

Der *gPAS* unterstützt, neben den in Tabelle 3 genannten Anforderungen, die Vergabe domänenspezifischer Pseudonyme, d.h. einem Studienteilnehmer können beliebig viele Pseudonyme je nach Studienzentrum, Art der Daten (Formulardaten, Bioproben) oder Kontext (Datenerfassung, Datenherausgabe) zugeordnet werden. Auf diese Weise können beispielsweise unerlaubte Rückschlüsse auf die Zusammengehörigkeit von medizinischen Daten, basierend auf dem verwendeten Pseudonym, verhindert werden.

Der *gICS* erlaubt die einfache und gleichzeitig tagesaktuelle Prüfung des Einwilligungstatus bereits erhobener medizinischer Daten, z.B. für die Datenherausgabe durch eine Transferstelle. Ein manuelles und zeitintensives Prüfen von Unterlagen entfällt dadurch. Die Einwilligungen werden dabei *gICS*-intern in Form von Policies (z.B. zur Datenherausgabe im Rahmen der Sekundärnutzung von Forschungsdaten) und Modulen (gruppierte Policies) abgebildet. Die Verwendung von versionierten Einwilligungsvorlagen (Templates) gestattet

dem Proband zu ausgewählten Modulen einzuwilligen. Auf diese Weise können auch komplexe Sachverhalte abgebildet und automatisiert geprüft werden (z.B. Zustimmung zur Datenerhebung bei gleichzeitiger Verweigerung der Entnahme von Bioproben) [38].

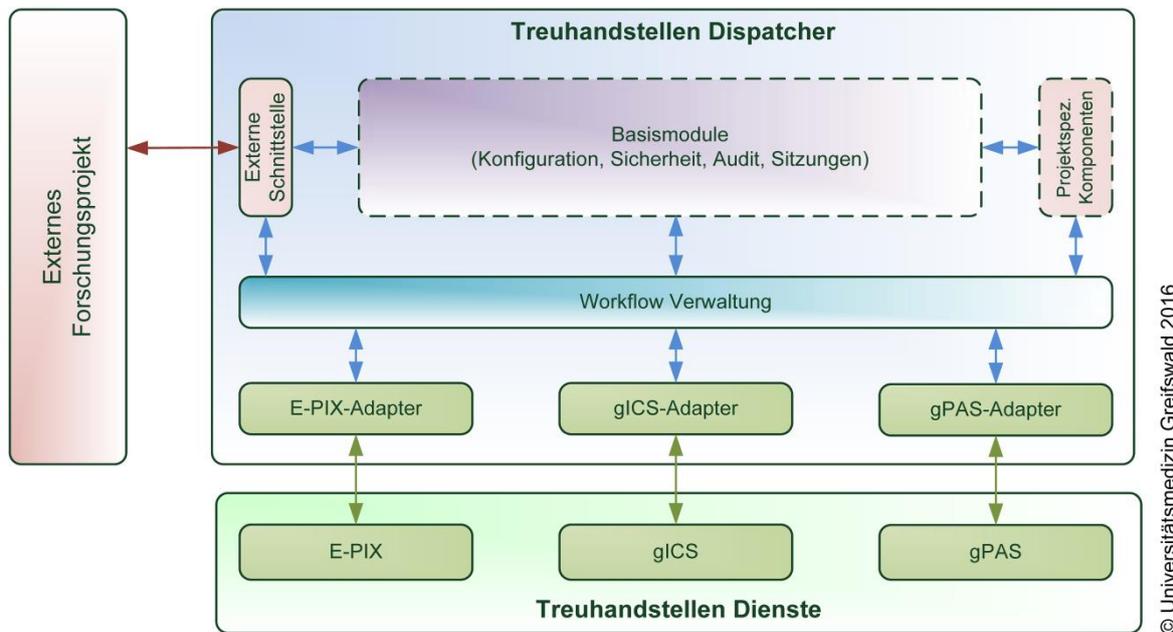
Jedes der Werkzeuge stellt dem Datentreuhänder eine eigenständige, grafische Benutzeroberfläche bereit. Ein direkter Informationsaustausch zwischen den einzelnen Werkzeugen ist, ohne den Einsatz zusätzlicher Software, nicht angedacht.

### **Übertragbarkeit des Treuhandstellenansatzes**

Die Charakteristika einer Treuhandstellenlösung sind abhängig von dem Szenario des Forschungsprojekts. Zum Beispiel erfolgt die Datenerhebung im *Zentralen klinischen Krebsregister MV* (ZKKR) [39] auf gesetzlicher Grundlage (Meldepflicht mit Widerspruchsmöglichkeit), dies macht ein Einwilligungsverfahren im oben beschriebenen Verfahren obsolet. Im Setting der Nationalen Kohorte wird aufgrund sehr umfangreicher Daten-, Bild- und Biomaterialerhebungen eine besondere Vielzahl unterschiedlicher Pseudonyme je Studienstandort und Datentyp (MRT, Bioproben, Formulardaten, etc.) generiert [36].

Jedes Szenario macht eine individuelle Anwendung der Werkzeuge E-PIX, gPAS und gICS in Bezug auf deren Konfiguration und interne Arbeitsabläufe der Treuhandstelle notwendig. Um interne Prozesse entsprechend automatisieren und die dafür erforderliche Kommunikation mit den prinzipiell voneinander unabhängigen Software-Modulen E-PIX, gPAS und gICS koordinieren zu können, wird ein Treuhandstellen-Dispatcher eingesetzt.

Die Funktionalität des Dispatchers kann über eine externe Schnittstelle zur Nutzung durch autorisierte Partner (z.B. Studienstandorte) zentral bereitgestellt werden. Der Dispatcher selbst hält zahlreiche Basismodule bereit, die u.a. die Konfiguration, die Sicherheit (Rollen und Rechte) sowie die Prozess-Auditierung unterstützen. Wesentliche Vorteile des Dispatchers sind *Interoperabilität* und *Erweiterbarkeit*. Der Treuhandstellen-Dispatcher greift nur indirekt auf die Funktionalität von E-PIX, gPAS und gICS über entsprechende Adapter zu. Auf diese Weise sind beliebige weitere Werkzeuge in das System integrierbar. Gleichzeitig kann das System je nach Bedarf durch projektspezifische Komponenten erweitert werden. Abbildung 4 zeigt zusammenfassend die grundlegende Architektur der Treuhandstelle unter Einsatz des Dispatchers (nach [32]).



© Universitätsmedizin Greifswald 2016

Abbildung 4 Beispiel für eine Treuhandstellenarchitektur (nach [32])

Um der *Individualität* eines Projekt szenarios entsprechen zu können und ebenso flexibel die erforderlichen Prozesse einer weiteren Kohortenstudie bzw. eines Registers durch neue Funktionalitäten im Dispatcher abbilden zu können, wurde ein workflow-orientierter Ansatz realisiert.

Workflows stellen den Kern der Individualisierbarkeit des Systems dar. Ein Workflow kann als Folge paralleler (oder auch sequentieller) Prozesse und Operationen verstanden werden und wird für die Steuerung der erforderlichen Funktionsaufrufe an die zugeordneten Softwaremodule (exemplarisch: E-PIX, gPAS und gICS) verwendet.

Um *Nachhaltigkeit* und *Nachnutzbarkeit* des Ansatzes zu unterstützen erfolgt die Unterscheidung in Basis-Workflows und projektspezifische Workflows. Basis-Workflows entsprechen typischen Arbeitsabläufen eines Datentreuhänders, wogegen projektspezifische Workflows die individuellen Charakteristika einer Studie bzw. eines Registers abbilden.

Der Treuhandstellen-Dispatcher verfügt bereits über eine umfangreiche Auswahl von Basis-Workflows (u.a. für das Anlegen eines Patienten mittels E-PIX, das Generieren eines Pseudonyms mittels gPAS oder auch das Überprüfen von Einwilligungen mittels gICS). Diese können als Basis für die Erstellung projektspezifischer Workflows genutzt werden [32] (vgl. Tabelle 5).

Workflow	Beschreibung
<b>get_mpi</b>	Generierung einer MPI ID für spezifizierte IDAT mittels E-PIX.
<b>check_patient_exists</b>	Prüfung, ob ein Teilnehmer mit den angegebenen IDAT bereits in der Datenbank des E-PIX existiert.
<b>get_id_from_id</b>	Anforderung eines Pseudonyms für eine angegebene Kennung (z.B. MPI ID) bzw. umgekehrt unter Verwendung des gPAS.
<b>add_consent</b>	Ablage eines IC (auf Grundlage zuvor spezifizierten Templates aus Modulen und Policies) für angegebene Kennung mittels gICS.
<b>check_consent_exists</b>	Prüfung, ob eine gültige Einwilligung für die angegebene Kennung in der gICS-Datenbank existiert.
<b>query_consent</b>	Abfrage einer Liste von Policies und von deren Einwilligungsstatus anhand einer zuvor vergebenen Informed Consent Kennung unter Verwendung des gICS.
<b>add_scan</b>	Anhängen eines digitalen Dokument-Scans an einen zuvor angelegten IC mittels gICS.
<b>update_participant</b>	Aktualisierung von IDAT eines Teilnehmers, der bereits in der E-PIX Datenbank angelegt wurde.
<b>get_participant_by_mpi_id</b>	Anforderung von IDATs eines Teilnehmers aus der E-PIX Datenbank anhand der MPI ID.
<b>add_participant_get_psn</b>	Sequentieller Workflow: Kombination von get_mpi und get_id_from_id
<b>get_participant_by_psn</b>	Sequentieller Workflow: Kombination von get_id_from_id und get_participant_by_mpi_id.

**Tabelle 5 Übersicht grundlegender Treuhandstellen-Workflows (nach [32])**

### 2.3 Evaluation der Praxistauglichkeit

Die Evaluation der Praxistauglichkeit des werkzeuggestützten Ansatzes zur Realisierung eines zentralen Datenmanagements soll nachfolgend exemplarisch für die Treuhandstellenwerkzeuge E-PIX, gPAS und gICS untersucht werden.

#### Bestimmung des Untersuchungsverfahrens

Nach ISO 9241 ist die Gebrauchstauglichkeit eines Produktes bestimmt durch das Ausmaß, „in dem [es] durch bestimmte Benutzer in einem bestimmten Nutzungskontext genutzt werden kann, um bestimmte Ziele effektiv, effizient und mit Zufriedenheit zu erreichen“ [40].

E-PIX, gPAS und gICS wurden mit Fokus auf größtmögliche Flexibilität entwickelt und sollen in kleinen und großen Projekten der epidemiologischen Forschung zum Einsatz kommen. In Anlehnung an ISO 9241 können diese Werkzeuge als praxistauglich betrachtet werden, wenn sie in möglichst diversen Forschungsprojekten unterschiedlicher Größe genutzt werden und gleichzeitig den vorgesehenen Zweck (Identitätsmanagement, Generierung und Verwalten von Pseudonymen, Einwilligungsmanagement) erfüllen. Um die Anwendung der Werkzeuge in der wissenschaftlichen Gemeinschaft zu untersuchen, kommen unterschiedliche Methoden in Frage.

#### Option A - Verfolgung der Nutzeraktivität

Die ausgewählten Werkzeuge werden derzeit einheitlich über die Projekt-Homepage *mosaic-greifswald.de* bereitgestellt (DFG-Förderkennzeichen HO 1937/2-1). Um die Werkzeuge herunterzuladen, ist eine einmalige Registrierung erforderlich. Beides ist aus rein technischer Sicht als Grundlage zum Anlegen einer Downloadstatistik und dessen Auswertung geeignet.

Jedoch verfügt diese Vorgehensweise über zu wenig Aussagekraft. Vom Download eines Werkzeugs kann nicht, ohne Zuhilfenahme weiterer Methoden, darauf geschlossen werden, ob und in welchem Umfang das Werkzeug tatsächlich in einem Projekt genutzt wurde.

#### *Option B – Strukturiertes Interview*

Durch die erforderliche Registrierung zum Download der Werkzeuge stehen konkrete Informationen für die direkte Kontaktaufnahme mit potentiellen Nutzern (u.a. Vor- und Nachname, Telefon, Mail-Adresse) zur Verfügung. In einem persönlichen Interview (Telefonat) könnten Anwender also direkt um Feedback in Bezug auf die Praxistauglichkeit der Werkzeuge gebeten werden und dieses als Grundlage der Auswertung genutzt werden. Die Bewertung der Werkzeuge und deren Funktionen erfolgt in diesem Fall auf Basis gesammelter Eindrücke und Erfahrungen. Laut Eden [41] ist diese Vorgehensweise allerdings zu wenig objektiv. Denn auch in einem stark strukturierten Interview kann der Nutzer selbst über Anzahl, Folge und Umfang seiner Qualitätsaussagen entscheiden. Dies kann den Verlust von Systematik zur Folge haben und sich dadurch negativ auf die methodische Qualität des Interviews auswirken. [41]

#### *Option C - Nutzerbefragung mittels Online-Fragebogen*

Eine abschließende Befragung der Werkzeug-Anwender mittels E-Mail Kontaktaufnahme und Online-Fragebogen (unter Berücksichtigung von DIN ISO 9241 [40]) erlaubt es, Akzeptanzaussagen von Nutzern auf systematische Weise zu ermitteln. Jedoch ist die Bereitschaft von Anwendern zur Teilnahme an solchen Online-Surveys im Allgemeinen gering. Erfahrungen einer Online-Umfrage-Plattform zeigen, dass E-Mail-basierte Umfragen vorwiegend nur eine geringe Beteiligung aufweisen (durchschnittlich rund 25 % Response Rate) [42]. Dies kann zu Verzerrungen führen (z.B. Selektionsbias), die sich negativ auf die Aussagekraft der Ergebnisse auswirken.

#### *Option D - Kontinuierliche Erhebung von werkzeugspezifischen Kennzahlen mit Projektbezug*

Eine weitere Option stellt die projekt- und gleichzeitig werkzeugspezifische Erhebung von Kennzahlen dar. Die zyklische Kennzahlenerhebung legt den Fokus der Untersuchung auf das Anwenderprojekt und ermöglicht die Nutzung der eingesetzten Werkzeuge sowohl quantitativ als auch objektiv zu betrachten. Die zu untersuchenden Merkmale können frei definiert werden, wie u.a. die Anzahl eingeschlossener Patienten, erzeugter Pseudonyme oder auch unterzeichneter Einwilligungen. Die Erhebung der Kennzahlen kann problemlos

durch Anpassung der Werkzeuge realisiert und durch entsprechende interaktive Kommunikationskanäle mit den Nutzern unterstützt werden (Telefon, E-Mail).

Für die Bestimmung der Praxistauglichkeit und Etablierung eines Werkzeugs in der wissenschaftlichen Gemeinschaft kann zwar die Anzahl der nutzenden Projekte betrachtet werden, wesentlich mehr Aussagekraft bietet jedoch die quantitative Erhebung ausgewählter Werkzeugmerkmale im spezifischen Projektkontext. Die Erhebung von Kennzahlen (Option D) bietet die Möglichkeit, Merkmale in Bezug auf die Werkzeuge E-PIX, gPAS und gICS zu definieren und deren Erhebung in strukturierter und gleichzeitig objektiver Weise durchzuführen.

### **Durchführung der Kennzahlenerhebung**

Um die Kennzahlen einheitlich und möglichst automatisiert ermitteln zu können, wurden die den Werkzeugen E-PIX, gPAS und gICS zugrundeliegenden MySQL-Datenbanken um entsprechend flexible Tabellen und direkt aufrufbare Prozeduren zur Aktualisierung der Kennzahlen erweitert (auf diese Weise können im Bedarfsfall nachträglich weitere Merkmale ergänzt werden). Beispielsweise werden durch Aufruf der neuen MySQL-Prozedur *updateStats* werkzeugspezifische Kennzahlen generiert, mit einem Zeitstempel versehen und in einer separaten Datenbanktabelle historisiert. Der tatsächliche Datenbestand von E-PIX, gPAS und gICS bleibt davon unberührt.

Um diese neue Funktionalität zur Nutzung verfügbar zu machen, wurden die über MOSAIC bereitgestellten Installationskripte für neue Projekte weiterentwickelt. Bestandsprojekten, die eines der Werkzeuge bereits vor Integration der Kennzahlenfunktionalität nutzten, wurden entsprechende Anpassungskripte und Hinweise zur Nutzung zur Verfügung gestellt. Von Januar 2015 bis Mai 2016 wurde die Erhebung der Kennzahlen in etwa 2-monatigem Abstand durchgeführt. Die Ansprechpartner in den Anwenderprojekten wurden per Mail über den Erhebungszeitpunkt informiert und gebeten, die automatisch generierten Werte im CSV-Format zu übermitteln. Diese Informationen wurden im Anschluss in eine zentrale Kennzahlendatenbank integriert. Dies erlaubt eine werkzeugspezifische, projektspezifische oder auch verlaufsspezifische Zusammenstellung der Daten zum Zweck der Auswertung.

Zusammenfassend zeigt Tabelle 6 welche Werkzeugmerkmale jeweils in Form von Kennzahlen im Detail erhoben wurden.

<b>E-PIX</b>	<b>gPAS</b>	<b>gICS</b>
Personen Unbearbeitete Matches Getrennte Matches Zusammengeführte Matches	Pseudonyme Anonyme Pseudonymdomänen	Einwilligungen Widerrufe Policies Module Templates Unterzeichnete Policies

**Tabelle 6 Erhobene Kennzahlen (Werkzeugmerkmale) der Werkzeuge E-PIX, gPAS und gICS**

Die entsprechenden Evaluierungsdimensionen (Auswahl der Werkzeugmerkmale) werden durch den Anwendungszweck der einzelnen Werkzeuge vorgegeben. Die Kennzahlen des E-PIX geben Aufschluss über das Verhältnis eingeschlossener Personen und erkannter möglicher Doppler (Matches). Ebenso wird der Bearbeitungsstand zur Bereinigung dieser möglichen Doppler deutlich. Neben der absoluten Anzahl verwalteter Pseudonyme zeigen die Kennzahlen des gPAS u.a. den Bedarf studienspezifischer Mehrfachpseudonymisierung (Anzahl der Domänen). Die Kennzahlen des gICS hingegen machen das Verhältnis von unterzeichneten Einwilligungen und eingegangenen Widerrufen deutlich. Zudem kann die Granularität der Einwilligungen (Verhältnis von Policies, Modulen und Templates), aber auch die Komplexität des Einwilligungsprozesses betrachtet werden.

Primär soll mit Hilfe der Kennzahlen projektübergreifend analysiert werden, wie viele Personen, Pseudonyme und Einwilligungen verwaltet werden. Zusätzlich sind weiterführende Analysen u.a. in Bezug auf zeitliche Verläufe, aber auch die kontextübergreifende Anwendung der Werkzeuge (gleichzeitige Anwendung von E-PIX, gPAS und gICS) in unterschiedlichen Anwendungsszenarien bzw. im Projektvergleich denkbar.

### 3 Ergebnisse

Schwerpunkt dieser Arbeit war es, wesentliche Komponenten eines zentralen Datenmanagements zu ermitteln und geeignete Werkzeuge für dessen Realisierung zu identifizieren. Im Anschluss sollte dargestellt werden, inwieweit der Aufbau einer Treuhandstelle im Rahmen eines zentralen Datenmanagements durch ausgewählte Werkzeuge unterstützt werden kann. Abschließend war zu evaluieren, wie dieser werkzeuggestützte Ansatz konkret in der epidemiologischen Praxis genutzt wird.

#### 3.1 Kernkomponenten des zentralen Datenmanagements

Auf Basis ausgewählter epidemiologischer Projekte und fachbezogener Publikationen wurden wesentliche funktionale und nicht-funktionale Anforderungen an ein zentrales Datenmanagement ermittelt und zusammengefasst (vgl. Tabelle 1). Darauf aufbauend wurden wesentliche Elemente eines ZDM abgeleitet und deren Zusammenspiel in einer beispielhaften Architektur veranschaulicht (vgl. Abbildung 2). Kernkomponenten sind Datenquelle, ETL-Prozess, Treuhandstelle, Speicherlösung und Datenbereitstellungsverfahren (Use and Access). Diese variieren je Studiensetting in Form und Umfang, sind aber grundlegend stets Bestandteil eines ZDM.

Am Beispiel der Kernkomponente Treuhandstelle (insbesondere ID-Management, Pseudonymisierung/Anonymisierung, Einwilligungsmanagement) wurden existierende Werkzeuge untersucht (vgl. Tabelle 2 – 4). Zusammenfassend bieten nur die Werkzeuge E-PIX, gPAS und gICS den erforderlichen Funktionsumfang, um die Realisierung einer Treuhandstelle im Rahmen eines ZDM für Kohortenstudien und Register umfassend zu unterstützen. Die ID-Management-Lösung E-PIX verwaltet Personen (-identitäten), erkennt potentielle Dubletten und unterstützt bei deren Auflösung. Der Pseudonymisierungsdienst gPAS generiert frei konfigurierbare Pseudonyme auf der Basis von Zufallszahlen und hilft bei deren Verwaltung. Die Einwilligungsmanagement-Lösung gICS unterstützt die Verwaltung von modularen Einwilligungen und Widerrufen [10] im Studienkontext. Diese und andere Werkzeuge werden durch das DFG-geförderte Projekt MOSAIC (Fördernummer HO 1937/2-1) bereitgestellt. Ziel dieses Projektes war es, vor allem kleineren epidemiologischen Forschungsprojekten durch praxis-orientierte modulare Werkzeuge bei der Lösung typischer Herausforderungen beim Aufbau eines zentralen Datenmanagements konkret zu helfen. Über das Web-Portal *mosaic-greifswald.de* werden unterschiedliche Werkzeuge und die zugehörige Dokumentation kostenfrei zur Verfügung gestellt [10]. Neben zahlreichen direkt

nutzbaren Vorlagen (Erstellung Datenschutzkonzept, Realisierung Datensicherung), Leitfäden (Erstellung Data Dictionary und webbasierte Erhebungsbögen) und Empfehlungen, umfassen weitere Hilfsmittel u.a. eine „*Toolbox for Research*“, die vor allem kleineren Projekten bei der Realisierung einer webbasierten, föderierten Datenerhebung helfen soll und eine R-Skript-Bibliothek, mit der ein studienbegleitendes Daten-Monitoring und eine grundlegende Prüfung der Qualität der erhobenen Daten realisiert werden kann [43].

Sämtlichen durch MOSAIC bereitgestellten Werkzeugen liegt langjährige praktische Erfahrung im Bereich des Datenmanagements aus zahlreichen Projekten [4, 5, 36, 44] zugrunde. Im Rahmen der Bereitstellung durch MOSAIC wurden existierende Lösungsansätze entlang der abgestimmten Leitlinien der TMF [2] überarbeitet und daraus entstehende Werkzeuge hinsichtlich Benutzerfreundlichkeit, Nachnutzbarkeit und Dokumentation verbessert. [10]

### **3.2 Werkzeuggestützter Aufbau einer Treuhandstelle im Rahmen des zentralen Datenmanagements**

Aus technischer, organisatorischer und personeller Sicht stellt der Aufbau einer Treuhandstelle eine wesentliche Herausforderung bei der Realisierung eines ZDM dar. Um den hohen Anforderungen des Datenschutzes entsprechen zu können (vgl. 2.2), sind zentrale, technische Verantwortlichkeiten einer Treuhandstelle (ID-Management, Pseudonymisierung, Einwilligungsmanagement) umzusetzen. Dies ist unter Zuhilfenahme der Werkzeuge E-PIX, gPAS und gICS möglich [32].

Unterschiede zwischen individuellen Kohortenstudien und Registern machen die Abbildung Szenario-spezifischer Abläufe innerhalb einer Treuhandstelle erforderlich. Diese lassen sich software-seitig und effektiv durch jeweils spezifische Kombination der Funktionalitäten der einzelnen Werkzeuge mittels eines Treuhandstellen-Dispatchers realisieren. Dabei gestattet ein workflow-basierter Ansatz erforderliche Anpassungen je Studie bzw. Register auf ein Mindestmaß zu reduzieren.

Bereits integrierte Workflows (vgl. Tabelle 5) lassen sich nahezu beliebig zu projektspezifischen Ausprägungen kombinieren, um individuellen Anforderungen gerecht zu werden. Auf diese Weise wird auch die Automatisierung von Prozessen gefördert (z.B. Anlegen einer Person (E-PIX) mit anschließender Generierung eines Pseudonyms (gPAS), automatischer digitaler Ablage der pseudonymisierten Einwilligung (gICS) und Übermittlung des Pseudonyms zur Datenerfassung an das Studienzentrum (vgl. Abbildung 5). Durch

Einsatz des Treuhandstellen-Dispatchers und entsprechender Workflows ist im gesamten Prozess keine manuelle Intervention des Datentreuhänders erforderlich. [32]

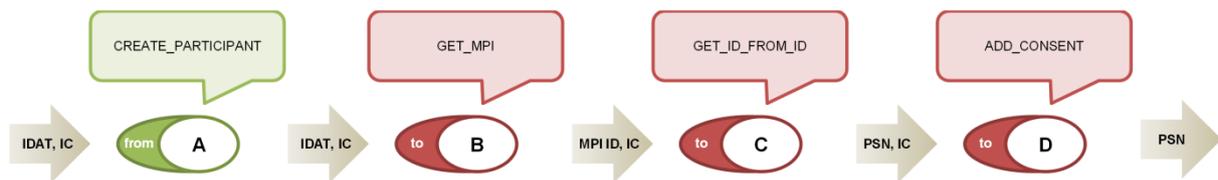


Abbildung 5 Die Abbildung eines Workflows am Beispiel eines Anwendungsfalls in der Treuhandstelle des Deutschen Zentrums für Herz-Kreislauf-Forschung e.V. (DZHK): personenidentifizierende Daten (IDAT) und die unterzeichnete informierte Einwilligung (IC) des Teilnehmers werden von der Treuhandstelle entgegengenommen, mittels ID-Management E-PIX, Pseudonymisierungswerkzeug gPAS und Einwilligungsmanagement gICS prozessiert und ein Pseudonym (PSN) zur Erfassung der medizinischen Daten (MDAT) an das Studienzentrum zurück übermittelt. [32]

### 3.3 Evaluation der Praxistauglichkeit

Laut Dumas et al. „besitzt ein Produkt selbst keinerlei Wert, es ist nur von Wert insofern es genutzt wird. Nutzung impliziert Nutzer“ [45]. Um also die Praxistauglichkeit eines werkzeuggestützten Ansatzes evaluieren zu können, wurde der Einsatz der MOSAIC-Werkzeuge in der Community (an dieser Stelle exemplarisch für E-PIX, gPAS und gICS) untersucht.

Seit 2015 wurde im Rahmen des MOSAIC-Projekts eine Kennzahlenerhebung (in Bezug auf die in Tabelle 6 gezeigten Merkmale) in allen bekannten Anwenderprojekten durchgeführt.<sup>5</sup> Tabelle 7 veranschaulicht die derzeitige Verbreitung der MOSAIC-Werkzeuge in Studien und Registern der epidemiologischen Forschung [43]. Alle Anwenderprojekte sind voneinander unabhängig. Sie unterscheiden sich wesentlich in den Themen der wissenschaftlichen Fragestellung, der geplanten Laufzeit, der Anzahl der einbezogenen Zentren, der Anzahl einzuschließender Teilnehmer und den eingesetzten Werkzeugen.

Um die Kennzahlenerhebung einheitlich durchführen zu können und die Merkmalsenerhebung zu vereinfachen, wurden E-PIX, gPAS und gICS um eine entsprechende Funktionalität erweitert. Die Abbildungen 6 – 8 zeigen die Verteilung der erhobenen Kennzahlen mit Projektbezug (die zugrundeliegenden Daten sind jeweils in den Tabellen 8 – 10 im Anhang dargestellt).

Die ausgewählten Werkzeuge werden in derzeit 8 abgeschlossenen und aktiven Projekten (vgl. Tabelle 7) eingesetzt. Dies zeigt, dass die Treuhandstellenwerkzeuge flexibel genug sind, um den unterschiedlichen Anforderungen sowohl kleinerer (z.B. MonDAFIS Studie der

<sup>5</sup> Die entsprechend der eingegangenen Rückmeldungen erfassten Kennzahlen werden im Anhang zusammenfassend in den Tabellen 8-10 dargestellt.

Charité Berlin) als auch größerer Forschungsprojekte (z.B. Nationale Kohorte, DZHK e.V.) zu entsprechen.

In Summe konnten mittels E-PIX bisher etwa 580.000 Personen erfasst, 2.5 Mio. Pseudonyme generiert und mittels gICS 69.000 Einwilligungen erfasst werden (Stand der Kennzahlenerhebung: 3. Mai 2016). Diese Ergebnisse der quantitativen Nutzungsanalyse belegen sowohl die Praxistauglichkeit der Treuhandstellenwerkzeuge als auch deren Nachnutzbarkeit in variierenden Studiensettings.

Status	Projekt	E-PIX	gPAS	gICS	Vorlage Datenschutz	Vorlage Datensicherung	Leitfaden Data Dictionary	Leitfaden eCRF	Toolbox for Research
<b>Aktiv</b>	Charité Schlaganfallzentrum MonDAFIS	x	x		x				
	Charité Schlaganfallzentrum Berliner Vorhofflimmern Register	x	x						
	Treuhandstelle des DZHK	x	x	x	x				
	DZHK Register TORCH				x				
	Treuhandstelle „ZDM“ der Medizinischen Fakultät der CAU zu Kiel	x	x	x	x				
	Summative Evaluation Kifög MV	x			x				
	Treuhandstelle des ZKKR-MV	x	x		x	x			
	Treuhandstelle des GANI_MED-Projekts	x	x	x					
	Treuhandstelle der Nationalen Kohorte	x	x	x	x				
	Nationales Verbrennungsregister		x				x	x	x
<b>In Vorbereitung</b>	Charité Schlaganfallzentrum - Treuhandstelle (CTMU)	x	x						
	DZHK Studie Transition				x				
	MVZ Labor Dessau GmbH	x	x						

**Tabelle 7 Überblick des Einsatzes der MOSAIC-Werkzeuge in der wissenschaftlichen Community [43]<sup>6</sup> (verwendete Abkürzungen: MonDAFIS = Impact of Standardized MONitoring for Detection of Atrial Fibrillation in Ischemic Stroke; DZHK = Deutsches Zentrum für Herz-Kreislauf-Forschung (DZHK) e.V.; TORCH=Translationales Register für Kardiomyopathien; CAU = Christian-Albrechts-Universität zu Kiel; Kifög MV = Summative Evaluation Kindertagesförderungsgesetz Mecklenburg Vorpommern; ZKKR-MV= Zentrales Klinisches Krebsregister MV; GANI\_MED = Greifswald Approach to Individualized Medicine; CTMU = vereinbarungsgemäße Struktureinheit der Charité; MVZ = Medizinisches Versorgungszentrum), Stand Januar 2016**

<sup>6</sup> Zum Zeitpunkt der letzten Erhebung liegen für das Projekt „Treuhandstelle ZDM der Medizinischen Fakultät der CAU zu Kiel“ noch keine Zahlen vor.

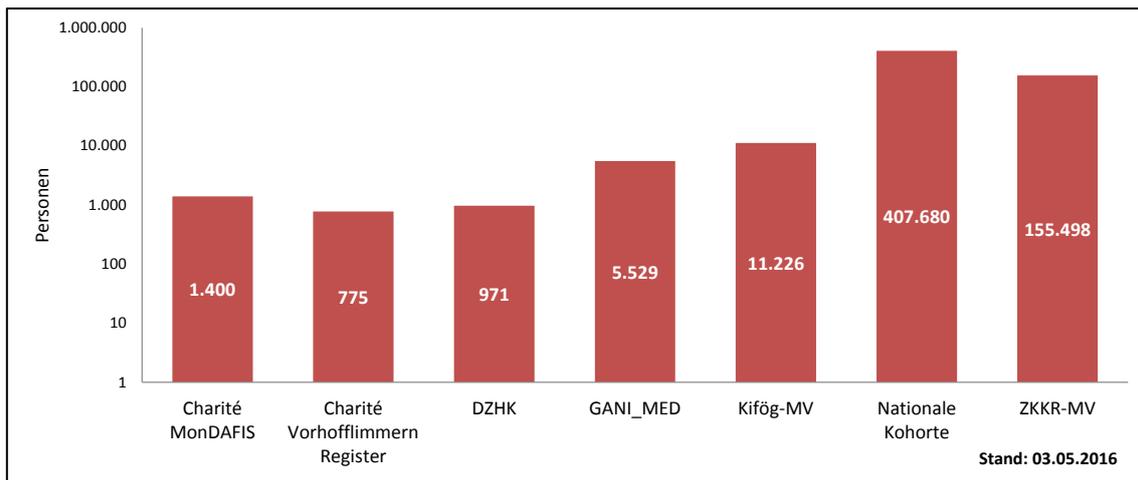


Abbildung 6 Anzahl der mittels E-PIX verwalteten Personen im Projektvergleich (583.079 Personen insgesamt, verwendete Abkürzungen gemäß Tabelle 7). Hinweis: Die Y-Achse wurde für die gewählte Darstellung logarithmisch transformiert.

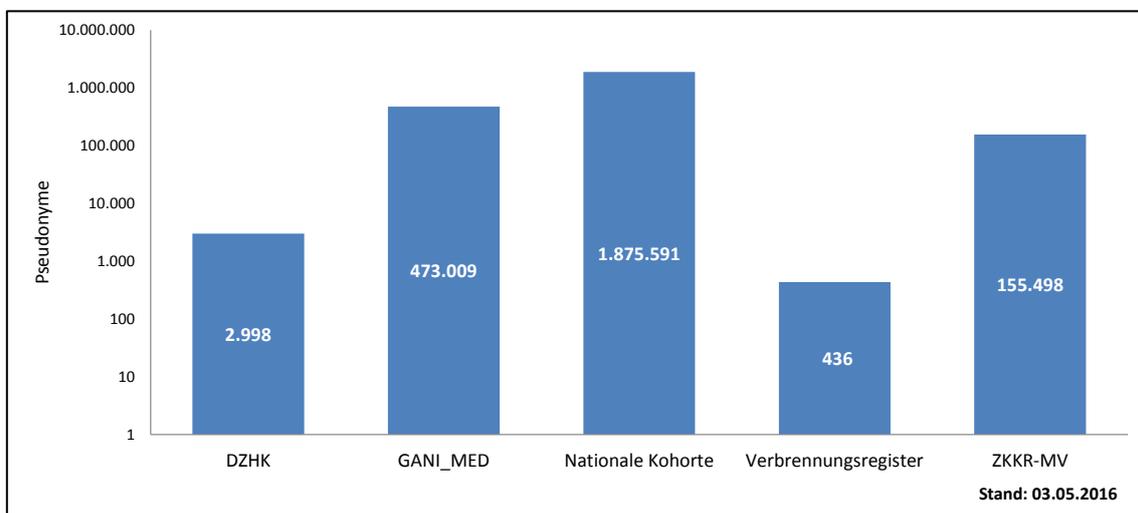


Abbildung 7 Anzahl der mittels gPAS generierten Pseudonyme im Projektvergleich (2.507.532 Pseudonyme insgesamt, verwendete Abkürzungen gemäß Tabelle 7). Hinweis: Die Y-Achse wurde für die gewählte Darstellung logarithmisch transformiert.

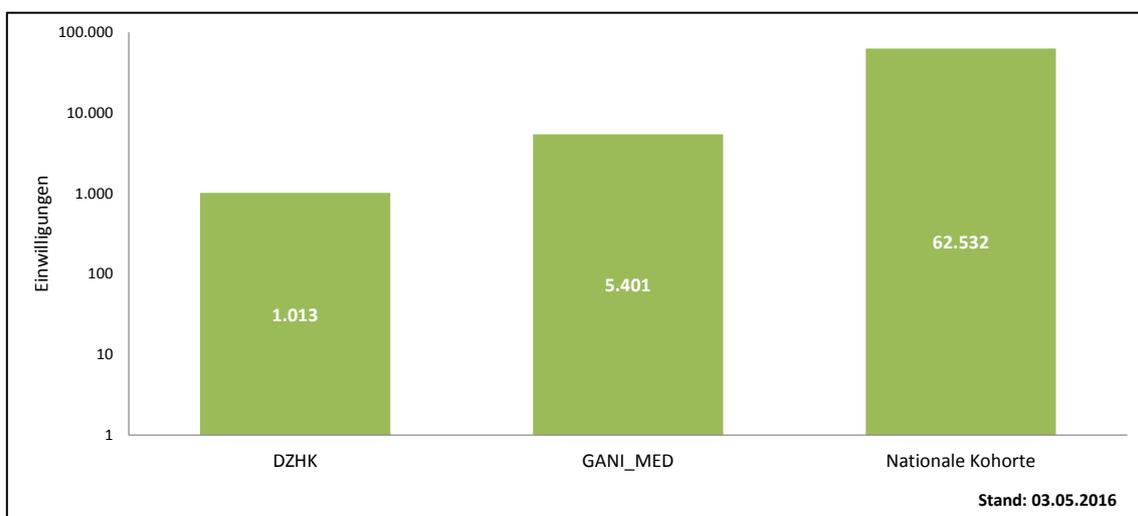


Abbildung 8 Anzahl der mittels gICS verwalteten Einwilligungen im Projektvergleich (68.946 Einwilligungen insgesamt, verwendete Abkürzungen gemäß Tabelle 7). Hinweis: Die Y-Achse wurde für die gewählte Darstellung logarithmisch transformiert.

## 4 Diskussion

In der wissenschaftlichen Community wurden in den letzten Jahren unterschiedliche Werkzeuge [10], aber auch umfangreiche IT-Plattformen (z.B. EHR4CR [46]) entwickelt. Aufgrund verschiedener Funktionsumfänge, unterschiedlicher technischer Rahmenbedingen, aber auch stark variierender Anwendungsszenarien war eine homogene Etablierung von ZDM Strategien in Kohortenstudien und Registern bislang nicht feststellbar.

Das durch die DFG geförderte Projekt MOSAIC stellt für ausgewählte Anforderungen eines zentralen Datenmanagements praxis-orientierte Unterstützung in Form von Vorlagen, Leitfäden und Werkzeugen kostenfrei bereit. Diese wurden gezielt für epidemiologische Forschungsprojekte mit geringen IT-Ressourcen verfügbar gemacht und können u.a. für den Aufbau einer unabhängigen Treuhandstelle, eine der Kernkomponenten des zentralen Datenmanagements, verwendet werden.

Dabei zeigen erfolgreich realisierte Treuhandstellen, beispielsweise im Rahmen des Deutschen Zentrums für Herz-Kreislauf-Forschung [44] oder der Nationalen Kohorte [36], dass der vorgestellte workflow-basierte Ansatz zum Aufbau einer Treuhandstelle die nötige Flexibilität bietet, unterschiedliche Studiendesigns und individuelle Datenschutzanforderungen eines Forschungsvorhabens unter Berücksichtigung des TMF Datenschutzleitfadens und in Übereinstimmung mit geltendem nationalen sowie internationalen Recht, abzubilden. [32]

Die Erhebung von Kennzahlen zeigte die Praxistauglichkeit eines werkzeugunterstützten Ansatzes zum Aufbau eines ZDM am Beispiel der Treuhandstellenwerkzeuge E-PIX, gPAS und gICS.

### 4.1 Stärken und Schwächen werkzeuggestützter Verfahren

Die durch MOSAIC bereitgestellten Werkzeuge und Vorlagen wurden mit Fokus auf Modularität, Übertragbarkeit und Nachnutzbarkeit entwickelt. Sie können unabhängig voneinander und durch einen allgemeingültigen (generischen) Ansatz in diversen Anwendungsfällen verwendet werden. Fraser et al. [47] haben gezeigt, dass durch Nachnutzung vorhandener Werkzeuge Kosten und Aufwände zur Durchführung von Studien gesenkt werden können.

Auch bei dem vorgestellten Treuhandstellenansatz wurde auf diese Stärke gesetzt. Der Treuhandstellen-Dispatcher unterstützt in besonderem Maße die Prozessautomatisierung. Dies unterstützt sowohl die Reduktion von Fehlern als auch die Steigerung der Effektivität.

Ein wesentlicher Schwachpunkt sind die derzeit noch erforderlichen Anpassungen und Konfigurationen, die im Rahmen der Einrichtung des Treuhandstellen-Dispatchers erforderlich werden. Dies ist bisher nur mit erheblichen IT-Ressourcen und Fachkenntnissen möglich. [32]

Die Ermittlung der Kennzahlen wurde auf Basis aller zurzeit bekannten Anwenderprojekte durchgeführt. Es ist nicht auszuschließen, dass weitere Anwender der Treuhandstellenwerkzeuge E-PIX, gPAS und gICS existieren, diese aber zum Erhebungszeitpunkt nicht bekannt waren oder aber bewusst sowohl Kontaktforderungen als auch initiale Kennzahlenanfragen nicht beantwortet haben. Auch ist die inhaltliche Validität und Vollständigkeit der erhaltenen Antworten nicht im Nachgang prüfbar.

## **4.2 Schlussfolgerungen**

Die Realisierung eines zentralen Datenmanagements, insbesondere der Aufbau einer Treuhandstelle für die datenschutzkonforme Verarbeitung personenbezogener Daten, stellt für eine Vielzahl epidemiologischer Forschungsprojekte eine erhebliche Herausforderung dar. Vor allem kleinere Vorhaben können die erforderlichen Aufwände oft nicht leisten, was zur Folge hat, dass bislang nur eine geringe Verbreitung zentralen Datenmanagements in Kohortenstudien und Registern feststellbar ist.

Die vorliegende Arbeit hat gezeigt, wie ein werkzeuggestützter Ansatz helfen kann, eine Vielzahl von Herausforderungen bei der Realisierung eines ZDM in Kohortenstudien und Registern zu meistern und gleichzeitig ressourcen- und zeitsparend dabei vorzugehen. Die damit einhergehende Reduktion von Nutzungshemmnissen kann sich positiv auf die Verbreitung von zentralem Datenmanagement, zumindest jedoch auf die Verbreitung der Treuhandstellenwerkzeuge E-PIX, gPAS und gICS, auswirken. Die ermittelten Kennzahlen legen nahe, dass dieser Ansatz in der epidemiologischen Praxis eine erhebliche Umsetzungsperspektive aufweist.

Die TMF und das MOSAIC-Projekt stellen eine Auswahl geeigneter Werkzeuge und Hilfsmittel bereit. Jedoch können auch im Idealfall diese Werkzeuge nicht alle Aspekte eines ZDM abdecken [10]. Beispiele sind der Betrieb und die Absicherung der notwendigen IT-Infrastrukturen, die Gewährleistung der technischen, organisatorischen und personellen Unabhängigkeit der Treuhandstelle oder auch die Prüfung des Datenschutzkonzeptes durch einen verantwortlichen Datenschutzbeauftragten des Landes bzw. Bundes [32]. Aus diesem Grund

können Unterstützungsangebote seitens der TMF oder MOSAIC in diesen Bereichen des ZDM bestenfalls anleitenden Charakter haben.

Die Qualität von Forschungsdaten und die Einhaltung datenschutzrechtlicher Rahmenbedingungen können durch Anwendung bereits erprobter Methoden gefördert werden [10].

Die untersuchten werkzeuggestützten Verfahren zur Realisierung einer Treuhandstelle tragen dazu bei, zentrales Datenmanagement in der epidemiologischen Forschung effektiver zu gestalten und somit die Forschung auf lange Sicht zu verbessern.

## 5 Zusammenfassung

Vor allem kleinere Forschungsvorhaben können die erforderlichen Aufwände zur Realisierung eines ZDM, insbesondere aber dem Aufbau einer Treuhandstelle, bislang häufig nicht leisten. Aufgrund vielzähliger Herausforderungen ist ZDM in Kohortenstudien und Registern daher nur wenig verbreitet [10].

Im Rahmen dieser Arbeit wurden, ausgehend von ausgewählten epidemiologischen Projekten und Fachpublikationen, wesentliche Anforderungen an ein ZDM zusammengefasst und zentrale funktionale Bestandteile eines ZDM identifiziert. Datenquellen, ETL-Prozesse, eine Treuhandstelle, eine Speicherlösung und ein Datenbereitstellungsverfahren sind Kernkomponenten eines ZDM. Am Beispiel der Treuhandstelle wurden erforderliche Werkzeuge identifiziert. Die ID-Management-Lösung E-PIX, das Pseudonymisierungswerkzeug gPAS und das Einwilligungsmanagement gICS bieten die notwendige Funktionalität. Alle werden kostenfrei über das MOSAIC-Projekt bereitgestellt.

Unterschiedliche Kohortenstudien und Register machen Szenario-spezifische Abläufe innerhalb einer Treuhandstelle erforderlich. Es wurde gezeigt, dass sich diese individuellen Abläufe software-seitig und effektiv durch Kombination der Funktionalitäten der einzelnen Werkzeuge (E-PIX, gPAS und gICS) in Form eines Treuhandstellen-Dispatchers realisieren lassen. Ein workflow-basierter Ansatz kann helfen, erforderliche individuelle Anpassungen auf ein Mindestmaß zu reduzieren.

Die Praxistauglichkeit dieses werkzeuggestützten Ansatzes wurde im Rahmen des MOSAIC-Projekts für die ausgewählten Werkzeuge mittels einer Kennzahlenerhebung in bekannten abgeschlossenen bzw. noch aktiven Anwenderprojekten (N=8) untersucht. In Summe konnten mittels E-PIX bisher etwa 580.000 Personen erfasst, 2.5 Mio. Pseudonyme generiert und mittels gICS 69.000 Einwilligungen erfasst werden (Stand: 03.05.2016). Weitere Anwendungen sind bereits in Vorbereitung. Der vorgestellte Treuhandstellenansatz wird bereits in zwei der Deutschen Zentren für Gesundheitsforschung genutzt [36, 44].

Auch wenn nicht jeder Aspekt eines ZDM durch vorkonfigurierte Werkzeuge unterstützt werden kann [10], wurde gezeigt, dass ein werkzeugunterstützter Ansatz zum Aufbau einer Treuhandstelle im Rahmen eines ZDM die nötige Flexibilität, Übertragbarkeit und Nachnutzbarkeit bietet, um den individuellen Anforderungen sowohl kleinerer als auch größerer Forschungsprojekte zu entsprechen und dabei gleichzeitig unterstützt, erforderliche Aufwände zu reduzieren.

## Literaturverzeichnis

- [1] Bundesamt für Sicherheit in der Informationstechnik, **BSI-Standard 100-2 IT-Grundschutz-Vorgehensweise**, Bonn, 2008.
- [2] K. Pommerening, J. Drepper, K. Helbing und T. Ganslandt, **Leitfaden zum Datenschutz in medizinischen Forschungsprojekten - Generische Lösungen der TMF – Version 2**, 1 Hrsg., Berlin, 2014.
- [3] O. C. Thamm, W. Perbix, J. Kricheldorf, R. Lefering, E. A. M. Neugebauer, B. Hartmann, B. Reichert und P. C. Fuchs, **„Etablierung eines nationalen Verbrennungsregisters (Abstract)“**, in *Deutsche Gesellschaft für Chirurgie. 131. Kongress der Deutschen Gesellschaft für Chirurgie. Berlin, 25.-28.03.2014.*, Düsseldorf, 2014.
- [4] H. Völzke, D. Alte, C. Schmidt, D. Radke, R. Lorbeer, N. Friedrich, N. Aumann, K. Lau, M. Piontek, G. Born, C. Havemann, T. Ittermann, S. Schipf, R. Haring, S. Baumeister, H. Wallaschofski, M. Nauck, S. Frick, A. Arnold, M. Jünger, J. Mayerle, M. Kraft, M. Lerch, M. Dörr, T. Reffelmann, K. Empen, S. Felix, A. Obst, B. Koch, S. Gläser, R. Ewert, I. Fietze, T. Penzel, M. Dören, W. Rathmann, J. Haerting, M. Hannemann, J. Röpcke, U. Schminke, C. Jürgens, F. Tost, R. Rettig, J. Kors, S. Ungerer, K. Hegenscheid, J. Kühn, J. Kühn, N. Hosten, R. Puls, J. Henke, O. Gloger, A. Teumer, G. Homuth, U. Völker, C. Schwahn, B. Holtfreter, I. Polzer, T. Kohlmann, H. Grabe, D. Roszkopf, H. Kroemer, T. Kocher, R. Biffar, U. John und W. Hoffmann, **„Cohort Profile: The Study of Health in Pomerania“**, *International Journal of Epidemiology*, Bd. 40, Nr. 2, S. 294-307, 4 2011.
- [5] H. J. Grabe, H. Assel, T. Bahls, M. Dörr, K. Endlich, N. Endlich, P. Erdmann, R. Ewert, S. B. Felix, B. Fiene, T. Fischer, S. Flessa, N. Friedrich, M. Gadebusch-Bondio, M. G. Salazar, E. Hammer, R. Haring, C. Havemann, M. Hecker, W. Hoffmann, B. Holtfreter, T. Kacprowski, K. Klein, T. Kocher, H. Kock, J. Krafczyk, J. Kuhn, M. Langanke, U. Lendeckel, M. M. Lerch, W. Lieb, R. Lorbeer, J. Mayerle, K. Meissner, H. M. zu Schwabedissen, M. Nauck, K. Ott, W. Rathmann, R. Rettig, C. Richardt, K. Saljé, U. Schminke, A. Schulz, M. Schwab, W. Siegmund, S. Stracke, K. Suhre, M. Ueffing, S. Ungerer, U. Völker, H. Völzke, H. Wallaschofski, V. Werner, M. T. Zygmunt und H. K. Kroemer, **„Cohort profile: Greifswald approach to individualized medicine (GANI\_MED)“**, *Journal of Translational Medicine*, Bd. 12, Nr. 144, 4 2014.
- [6] J. Meyer, S. Ostrzinski, D. Fredrich, C. Havemann, J. Krafczyk und W. Hoffmann, **„Efficient data management in a large-scale epidemiology research project“**, *Computer Methods and Programs in Biomedicine*, Bd. 107, Nr. 3, S. 425-35, 9 2012.
- [7] U. Jensen, **Leitlinien zum Management von Forschungsdaten**, GESIS - Leibniz-Institut für Sozialwissenschaften, Köln, 2012.

- [8] S. Higgins, „**The DCC Curation Lifecycle Model**,“ *The International Journal of Digital Curation*, Bd. 3, Nr. 1, 2008.
- [9] Universität Marburg, **Handlungsfelder von Forschungsdatenmanagement**, 2014. [Online]. URL: <https://www.uni-marburg.de/projekte/forschungsdaten/management/fodamanagen>. [Zugriff am 15 07 2014].
- [10] M. Bialke, T. Bahls, C. Havemann, J. Piegsa, K. Weitmann, T. Wegner und W. Hoffmann, „**MOSAIC. A modular approach to data management in epidemiological studies**,“ *METHODS OF INFORMATION IN MEDICINE*, Bd. 54, Nr. 4, S. 364-371, 8 2015.
- [11] U. John, E. Hensel, J. Lüdemann, M. Piek und S. Sauer, „**Study of Health in Pomerania (SHIP): A health examination survey in an east German region: Objectives and design**,“ *Sozial- und Präventivmedizin*, Bd. 46, Nr. 3, S. 186-94, 2001.
- [12] Nationale Kohorte e.V., **www.nationale-kohorte.de**, [Online]. URL: [http://www.nationale-kohorte.de/content/Datenschutzkonzept\\_130314.pdf](http://www.nationale-kohorte.de/content/Datenschutzkonzept_130314.pdf). [Zugriff am 24 10 2013].
- [13] C. Michalik, J. Dreß, J. Stausberg und S. M. N. Ngouongo, **Von der Evaluierung zur Konsolidierung: Anforderungen an Kohortenstudien & Register IT (KoRegIT) (Vortragsfolien)**, 25 9 2013. [Online]. URL: [http://www.tmf-ev.de/DesktopModules/Bring2mind/DMX/Download.aspx?Method=attachment&Command=Core\\_Download&EntryId=22278&PortalId=0](http://www.tmf-ev.de/DesktopModules/Bring2mind/DMX/Download.aspx?Method=attachment&Command=Core_Download&EntryId=22278&PortalId=0). [Zugriff am 22 10 2013].
- [14] M. Sariyar, A. Borg, O. Heidinger und K. Pommerening, „**A practical framework for data management processes and their evaluation in population based medical registries**,“ *Informatics for Health and Social care*, Bd. 38, Nr. 2, S. 104-19, 2013.
- [15] D. Müller, M. Augustin, N. Banik, W. Baumann, K. Bestehorn, J. Kieschke und R. Lefering, „**Memorandum Register für die Versorgungsforschung**,“ *Gesundheitswesen 2010*, Bd. 72, S. 824-839, 2010.
- [16] TMF e.V., **Rechtsgutachten zu Verwertungsfragen**, Berlin: Medizinisch Wissenschaftlichen Verlagsgesellschaft (MWV), 2008.
- [17] TMF e.V., **Informed Consent - Software wizard of the TMF provides practical support**, 2007. [Online]. URL: <http://www.tmf-ev.de/News/articleType/ArticleView/articleId/223.aspx>. [Zugriff am 29 01 2015].
- [18] TMF e.V., **TMF PID-Generator**, 2014. [Online]. URL: [http://www.tmf-ev.de/Themen/Projekte/V015\\_01\\_PID\\_Generator.aspx](http://www.tmf-ev.de/Themen/Projekte/V015_01_PID_Generator.aspx). [Zugriff am 10 02 2015].
- [19] TMF e.V., **TMF Pseudonymisierungsdienst (PSD)**, 2014. [Online]. URL: [http://www.tmf-ev.de/Projekte/TMFProjekte/V000\\_01\\_PSD.aspx](http://www.tmf-ev.de/Projekte/TMFProjekte/V000_01_PSD.aspx). [Zugriff am 16 02 2015].
- [20] TMF e.V., **TMF Metadata Repository**, 2014. [Online]. URL: [http://www.tmf-ev.de/Themen/Projekte/D021\\_01\\_Metadata\\_Repository.aspx](http://www.tmf-ev.de/Themen/Projekte/D021_01_Metadata_Repository.aspx). [Zugriff am 24 2 2015].

- [21] TMF e.V., **TMF Anon-Tool - Werkzeug zur Anonymisierung von Datenexporten**, 2013. [Online]. URL: [http://www.tmf-ev.de/Themen/Projekte/V08601\\_AnonTool.aspx](http://www.tmf-ev.de/Themen/Projekte/V08601_AnonTool.aspx). [Zugriff am 24 4 2013].
- [22] M. Rani und B. Buckley, „**Systematic archiving and access to health research data: rationale, current status and way forward**,“ *Bulletin of the World Health Organization*, Bd. 90, Nr. 12, S. 932-939, 12 2012.
- [23] Duden, **Duden Online ("Kern, der")**, [Online]. URL: <http://www.duden.de/rechtschreibung/Kern>. [Zugriff am 11 02 2016].
- [24] oose. Innovative Informatik, **Begriffserklärung Komponente**, [Online]. URL: <http://www.oose.de/glossar/komponente-2/>. [Zugriff am 11 02 2016].
- [25] M. Bialke, D. Langner, L. Geidel, T. Bahls, J. Piegsa, C. Havemann und W. Hoffmann, „**Who am I? And if so, how many?**“ – **The E-PIX as innovative system to manage person identities, Posterbeitrag, 2nd Data Management Workshop, Universität Köln**, 2014. [Online]. URL: <http://www.tr32db.uni-koeln.de/workshops/poster.php?wsID=4>. [Zugriff am 06 05 2016].
- [26] M. Lablans, A. Borg und F. Ückert, **unimedizin-mainz.de - Die Mainzliste als Open Source**, 2013. [Online]. URL: <http://www.unimedizin-mainz.de/imbei/informatik/opensource/mainzliste.html>. [Zugriff am 15 10 2013].
- [27] Sysnet International, **Open Enterprise Master Patient Index**, 2014. [Online]. URL: <http://www.openempi.org>. [Zugriff am 10 03 2014].
- [28] MOSAIC, **MOSAIC - ID-Management mittels E-PIX**, 2014. [Online]. URL: <https://mosaic-greifswald.de/werkzeuge-und-vorlagen/id-management-e-pix.html>. [Zugriff am 24 2 2015].
- [29] L. Geidel, T. Bahls und W. Hoffmann, „**Ein generisches Pseudonymisierungswerkzeug als Modul des Zentralen Datenmanagements medizinischer Forschungsdaten (Abstract)**,“ in *Abstractband 8. Jahrestagung der Deutschen Gesellschaft für Epidemiologie und 1. Internationales LIFE Symposium*, M. Löffler und S. Riedel-Heller, Hrsg., Leipzig, 2013, S. 245-246.
- [30] O. Heinze, M. Birkle, L. Köster und B. Bergh, „**Architecture of a consent management suite and integration into IHE-based regional health information networks**,“ *BMC Medical Informatics and Decision Making*, Bd. 11, Nr. 58, 10 2011.
- [31] T. Bahls, W. Liedtke, L. Geidel und M. Langanke, „**Ethics Meets IT: Aspects and elements of Computer-based informed consent processing**,“ in *Individualized medicine, ethical, economical and historical perspectives*, T. Fischer, M. Langanke, P. Marschall und S. Michl, Hrsg., Springer, 2015, S. 209-229.
- [32] M. Bialke, P. Penndorf, T. Wegner, T. Bahls, C. Havemann, J. Piegsa und W. Hoffmann, „**A workflow-driven approach to integrate generic software modules in a Trusted Third Party**,“ *Journal of Translational Medicine*, Bd. 13, Nr. 176, 6 2015.

- [33] Bundesrepublik Deutschland, **Bundesdatenschutzgesetz (BDSG) (Fassung vom 14.01.2003 (BGBl. I S. 66), letzte Änderung 25.02.2015 (BGBl. I S. 162) m.W.v. 01.01.2016)**, 2016.
- [34] Council of Europe, „**Convention for the protection of individuals with regard to automatic processing of personal data,**“ in *ETS No.108*, Strasbourg, 1981.
- [35] World Medical Association (WMA) (2008) WMA Declaration of Helsinki - Ethical Principles for Medical Research Involving Human Subjects, 59th WMA General Assembly Hrsg., Korea, 2008.
- [36] C. Havemann, K. Fitzer, S. Ostrzinski, R. Wolff, M. Bialke, T. Bahls und W. Hoffmann, **Datenschutz- und IT-Sicherheitskonzept für die unabhängige Treuhandstelle der Nationalen Kohorte**, 2014. [Online]. URL: <http://www.nationale-kohorte.de/content/treuhandstellenkonzept.pdf>. [Zugriff am 24.2.2015].
- [37] P. Schaar, „**Privacy by Design,**“ *Identity in the Information Society*, Bd. 3, Nr. 2, S. 267-274, 8.2010.
- [38] L. Geidel, T. Bahls und W. Hoffmann, „**Darf ich? – Herausforderungen an eine generische, automatisierte elektronische Verwaltung von Einwilligungen (Abstract),**“ in *GMDS 2014. 59. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (GMDS)*, Göttingen, 07.-10.09.2014, 2014.
- [39] U. Siewert, M. Freund, S. Scriba, V. Schreier, G. Dölken, G. Hilde und W. Hoffmann, „**Die Klinischen Krebsregister und das Zentrale Klinische Krebsregister Mecklenburg-Vorpommern: Gemeinsam für die Qualitätssicherung in der Onkologie.**“ *ÄRZTEBLATT MECKLENBURG-VORPOMMERN*, Nr. 2, S. 49-53, 2013.
- [40] International Organisation for Standardisation (Hrsg.), **ISO 9241 -Ergonomische Anforderungen für Büroarbeit mit Bildschirmgeräten Teil 1-17**, I. O. f. Standardisation, Hrsg., Berlin: Beuth Verlag, 1992-1997.
- [41] Jens Eden (Universität Oldenburg, Diplomarbeit), **Bewertung der Praxistauglichkeit von MUSE - Beurteilung der Gebrauchsmerkmale**, 1998. [Online]. URL: [http://www.cg-hci.informatik.uni-oldenburg.de/~da/eden/Inhalt/Kapitel\\_3/3.1.4beurteilung\\_der\\_gebrauchsmerkmale.htm#3.1.4](http://www.cg-hci.informatik.uni-oldenburg.de/~da/eden/Inhalt/Kapitel_3/3.1.4beurteilung_der_gebrauchsmerkmale.htm#3.1.4). [Zugriff am 10.03.2016].
- [42] Fluid Surveys University, **FluidSurveyUniversity: Response Rate Statistics for Online Surveys -What Numbers Should You be Aiming For?**, 2014. [Online]. URL: <http://fluidsurveys.com/university/response-rate-statistics-online-surveys-aiming/>. [Zugriff am 14.03.2016].
- [43] M. Bialke und W. Hoffmann, **Abschlussbericht zum DFG-Projekt "MOSAIC" (Modulares Unterstützungssystem zur Planung und Implementierung eines zentralen Datenmanagements für Forschungsprojekte der Gesundheitsforschung), LIS-Förderprogramm "Informationsinfrastrukturen"**, Greifswald, 2016.

- [44] German Centre for Cardiovascular Research (DZHK), **dzhk.de**, 11 03 2015. [Online]. URL: <http://dzhk.de/>. [Zugriff am 10 02 2015].
- [45] J. S. Dumas und J. C. Redish, **A Practical Guide to Usability Testing**, 2. Auflage, Rev. Ed. Hrsg., Intellect Books, 1999.
- [46] J. Doods, R. Bache, M. McGilchrist, C. Daniel, M. Dugas und F. Fritz, **„Piloting the EHR4CR Feasibility Platform across Europe,“** *Methods of Information in Medicine*, Bd. 53, Nr. 4, S. 264-268, 2014.
- [47] H. Fraser, D. Thomas, J. Tomaylla, N. Garcia, L. Lecca, M. Murray und M. Becerra, **„Adaptation of a web-based, open source electronic medical record system platform to support a large study of tuberculosis epidemiology,“** *BMC Medical Informatics and Decision Making*, Bd. 12, Nr. 125, 11 2012.
- [48] TMF e.V., **k-Anonymität und I-Diversität bieten sicheren Schutz vor dem Ausspionieren personenbezogener Daten (Interview mit Prof. Dr. Johann Eder über das neue „Anon“-Tool der TMF zur Anonymisierung medizinischer Daten)**, 3 2013. [Online]. URL: <http://www.tmf-ev.de/News/articleType/ArticleView/articleId/1270.aspx>. [Zugriff am 06 06 2016].

## Anhang

## A Wissenschaftliche Artikel

Original Articles

364

## MOSAIC – A Modular Approach to Data Management in Epidemiological Studies

M. Bialke<sup>1</sup>; T. Bahls<sup>1</sup>; C. Havemann<sup>1</sup>; J. Piegsa<sup>1</sup>; K. Weitmann<sup>1</sup>; T. Wegner<sup>2</sup>; W. Hoffmann<sup>1</sup>

<sup>1</sup>Institute for Community Medicine, Section Epidemiology of Health Care and Community Health, University Medicine Greifswald, Greifswald, Germany;

<sup>2</sup>Institute of Applied Microelectronics and Computer Engineering, University of Rostock, Rostock, Germany

### Keywords

Medical data management, data privacy protection, informed consent, pseudonyms, record linkage

### Summary

**Introduction:** In the context of an increasing number of multi-centric studies providing data from different sites and sources the necessity for central data management (CDM) becomes undeniable. This is exacerbated by a multiplicity of featured data types, formats and interfaces. In relation to methodological medical research the definition of central data management needs to be broadened beyond the simple storage and archiving of research data.

**Objectives:** This paper highlights typical requirements of CDM for cohort studies and registries and illustrates how orientation for CDM can be provided by addressing selected data management challenges.

**Methods:** Therefore in the first part of this paper a short review summarises technical, organisational and legal challenges for CDM in cohort studies and registries. A deduced set of typical requirements of CDM in epidemiological research follows.

**Results:** In the second part the MOSAIC project is introduced (a modular systematic approach to implement CDM). The modular nature of MOSAIC contributes to manage both technical and organisational challenges efficiently by providing practical tools. A short presentation of a first set of tools, aiming for selected CDM requirements in cohort studies and registries, comprises a template for comprehensive documentation of data protection measures, an interactive reference portal for gaining insights and sharing experiences, supplemented by modular software tools for generation and management of generic pseudonyms, for participant management and for sophisticated consent management.

**Conclusions:** Altogether, work within MOSAIC addresses existing challenges in epidemiological research in the context of CDM and facilitates the standardized collection of data with pre-programmed modules and provided document templates. The necessary effort for in-house programming is reduced, which accelerates the start of data collection.

### 1. Introduction

The collection and provision of medical data forms the basis for an analytical medical research. Therefore, cohort studies and registries mainly focus on quality-assured collection of primary research data and associated metadata as well as preservation of data interpretability.

When designing cohort studies and registries many research projects are confronted with considerable challenges (►Figure 1). Considering the phases of the data lifecycle [1], this concerns organisational and technical effort necessary for the realization of a comprehensive data management including collection, processing, long-term storage and provision of data pursuant to recommendations of accredited institutions (e.g. the DFG [2]). One particular challenge is the integration of data from heterogeneous source systems. This includes laboratory information and management systems (LIMS), clinical information systems (CIS), diagnostic equipment and electronic case report forms (eCRF) usually generating a wide variety of data types and formats (e.g. form or image data, CSV exports, GDT files, XML structures, HL7 messages). Consequently, suitable measures must be adopted in order to ensure data quality throughout processing activities, to enable automated validation and to support interactive data correction. In addition, compliance with legislation for data protection must be assured at both federal and state level. For this reason, the ethics model must be built around a central consent management system administrating individual items of consent and authorisation (i.e. informed consent) and permitting revocation cross-checking on a daily basis. Management of

### Correspondence to:

Martin Bialke  
Institute for Community Medicine  
Department Epidemiology of Health Care  
and Community Health  
University Medicine Greifswald  
Ellernholzstr. 1–2  
17487 Greifswald  
Germany  
E-mail: martin.bialke@uni-greifswald.de

Methods Inf Med 2015; 54: 364–371  
<http://dx.doi.org/10.3414/ME14-01-0133>  
received: December 5, 2014  
accepted: June 3, 2015  
epub ahead of print: July 21, 2015

participants is required in order to aggregate personal medical data within a central data repository while avoiding mistakes due to homonyms or synonyms. Furthermore, the ability to process and provide research data in a pseudonymised form is mandatory.

Therefore, the introduction of a central data management (CDM) aids to simplify the process of design and implementation of data management for cohort studies and registries. This is accomplished by providing structured, uniform and reproducible processes throughout all phases of the lifecycle of research data. Simultaneously, preconditions for long-term data usability and the comparability of derived results are constituted. More precisely, the use of CDM allows for the integration of all core issues for data management into cohort studies and registries as early as the planning stage [3]. In order to reduce effort, the deployment of reusable, modular solutions is encouraged. Thereby, the high effort arising from the repeated procedure of developing concepts and implementing dedicated software solutions for individual projects is avoided.

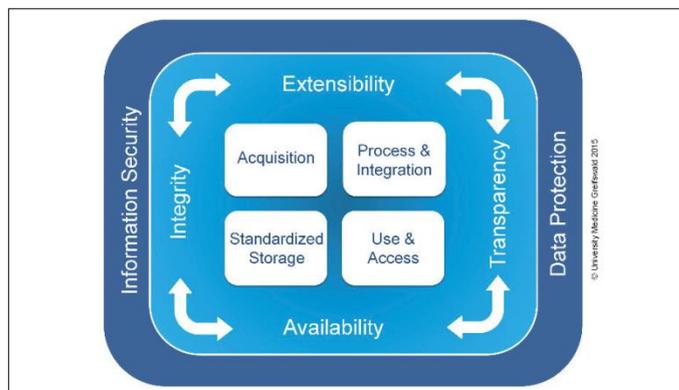
## 2. Objectives

Aim of this paper is to highlight typical requirements of CDM for cohort studies and registries, deduced from legal aspects, organisational task areas and technical approaches of data management typically applied within the phases of epidemiological projects. Based on cohort studies and registries, the paper illustrates how MOSAIC provides orientation for CDM by addressing selected data management challenges with a first set of ready-to-use and open source tools. MOSAIC is introduced to support the implementation of CDM for cohort studies and registries and to facilitate reducing in-house programming effort as well as promoting sustainability of existing solutions.

## 3. Methods

### 3.1 Central Data Management in Epidemiological Research

In cohort studies and registries many organisational and technical tasks must be completed prior to actual operation. ► Fig-



**Figure 1** CDM in the context of epidemiological cohort studies and registries: CDM comprises all technical and organisational measures aiming for data acquisition, data processing and integration, the standardized storage as well as the use and access of data. At the same time CDM has to fulfill non-functional requirements and relating interdependencies, such as requirements for data integrity, data availability, process transparency and system extensibility. Also CDM has to ensure conformity to legal standards of information security and data protection.

ure 2 depicts both organisational and a technical measures and parameters that need to be considered during the preparation phase and the subsequent phases of acquisition and usage. Starting with the specification of research questions and variables to be surveyed, the ethical framework (including determination of patient details and informed consent (e.g. paper-based or as digital document) for the study must be established first (**preparatory phase**). Additionally, appropriate strategies for implementing authorisations and revocations, and the definition of workflows to support these activities are determined. In order to guarantee the consistent participant management and the necessary privacy an unambiguous identification of persons and the pseudonymisation of personal data must be ensured [4]. Subsequently, the necessary techniques for data collection, processing, archiving and provisioning must be defined. Furthermore, the technical environment necessary to ensure protection of data privacy and information security must be specified. Specifications must also be drawn up for separation and storage of personal data and medical data. This comprises a data model capable of mapping the collected information, a data repository for information storage and a

role-based management system for access authorisation, to name but a few. Moreover, a data protection concept must be specified considering organisational, technical and personnel-related issues. This concept requires review and approval by the responsible data protection officer.

For the collection of research data, appropriate electronic case report forms (eCRFs) must be designed reverting to standard data collection tools including the data validation and possible data correction (automated and interactive). Amongst others, required interfaces, expected data formats, necessary metadata and data transfer protocols must be defined for data extraction from diagnostic equipment and other sources. In addition, measures to support record linkage, in order to minimize synonym and homonym errors during the data merging process [5], as well as to secure the data transfer between the study centres and the target storage system require specification.

During the subsequent **acquisition phase**, the focus is on ensuring high quality of the collected research data and on guaranteeing its security and long-term provisioning. Related tasks include the unobstructed operation of all systems required for error-free data collection as well as the definition and implementation of quality

assurance measures (e.g. instructing study personnel or source data checking and correction). At the same time, a high level of data integrity needs to be secured (i.e. by matching procedures for aggregation of personal data from multiple sources). Mechanisms for historisation and version control – aided by continuous monitoring and documentation of all data-handling processes – guarantee the necessary auditability of the data processing systems. In addition, backup strategies and rules regulating data access guarantee data security during this phase. To facilitate follow-up data collection, CDM also supports the re-contacting of study participants.

During the **usage phase** the collected medical data is provided to researchers while complying with stated requirements for data privacy and protection. This comprises well-defined procedures for use and access and study-specific pseudonymisation. Furthermore, free specification of variable sets and export formats is provided to researchers. A preferably automated procedure is applied to determine and request the associated pseudonymised data, following prior verification of the necessary consent. Subsequently, the requested data is transmitted via a dedicated transfer unit. On completion of a study, a

similar approach must be used to re-integrate the research results into the data pool. This includes derived variables, generated scores or variable coding work, for example. In the case of subsequent consent revocation or incidental findings subject to reporting obligations, the system must be capable of re-establishing the link from the data to individuals by a defined de-pseudonymisation procedure.

In summary, the study lifecycle requires a wide range of processes and measures that are not limited to separate phases within an individual study. On account of the considerable technical effort required to realize the necessary systems and functionalities, implementation should commence at the earliest possible juncture. However, the system architecture must exhibit sufficient versatility to satisfy supplementary requirements, changes and extensions in the study design.

Epidemiological studies and registries show that individual research projects exhibit particular requirements and settings [6–9]. Nevertheless, implementing a data management system as part of a study generates a recurrent set of similar issues. Such issues can be categorized to functional aspects, non-functional requirements and establishing compliance with legal frame-

works. Fundamentally, CDM is targeted on satisfying each of these requirements.

Meyer et al. [10] define non-functional requirements for a CDM system. The authors are particularly focussing on data management processes. Reverting to this, their implementation is described by suitable technical measures with focus on quality assurance concentrating on the processed data.

In the context of epidemiological cohort studies these technical, organisational and staffing measures required to implement data management can be grouped into primary task areas. Combined with the non-functional requirements, they constitute the most important core elements of CDM. Details on the individual measures that each core element typically involves are provided in ► Table 1.

Since the primary task of CDM is to aggregate data from heterogeneous and federated sources and to transform these data into suitable structures, the focus is on ensuring that the collection, processing, storage and provisioning of data is standardized and homogenized and coordinated from a single central point.

Therefore, the successful implementation of CDM in an epidemiological context is conditional on expert support (i.e. by com-

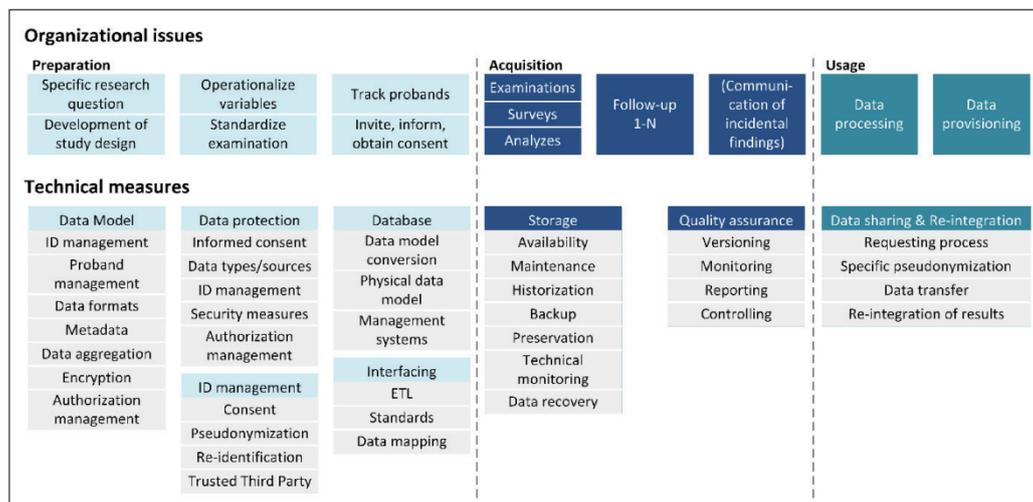


Figure 2 Typical phases of a cohort study: Organisational issues and technical measures concerning CDM regarding the phases of a cohort study

puter scientists and IT personnel) for the research. In addition, data management must be included in study planning at an early stage. Ideally, implementation of the technical and organisational requirements should be completed before data collection starts.

## 4. Results

### 4.1 A Modular Approach to Support the Implementation of CDM

In the context of the MOSAIC project (DFG program “Information infrastructure for research data”) a set of mutually independent tools is developed, each of them addressing a practical requirement of data management in cohort studies and registries (► Figure 2).

Primarily targeting newly initiated cohort studies and registries, the MOSAIC web-portal ([mosaic-greifswald.de](http://mosaic-greifswald.de)) presents guidance and insights to legal, organisational and technical aspects of CDM. Scientists’ attention is drawn to relevant literature, existing solutions (provided and recommended e.g. by the TMF [12]) and current issues, assisting and encouraging them to specify the individual requirements for their study setting. Addressing

these requirements and in accordance to the rules the TMF consented upon with the German protection officer [12], MOSAIC provides a portfolio of independent software modules, templates, checklists and recommendations in order to facilitate reuse of existing solutions. For already established cohort studies and registries, some MOSAIC tools might be of interest as well, e.g. if previously defined requirements have changed and new technical ways for participant management or pseudonymisation of existent medical data have to be identified.

The aim of MOSAIC is to provide support for the responsibilities of CDM through modular solutions (► Figure 3). This approach allows for focusing on selected challenges instead of establishing complete solutions for a data management lacking adaptability, interoperability and flexibility.

A particular benefit of modularity is enhanced re-usability. The tools might be utilized separately in one epidemiological project to address a specific issue or used in combination in another project to satisfy a set of requirements. Nevertheless, MOSAIC does not intend to provide a complete software suite serving as a stand-

alone solution for CDM. Each tool is or will be developed based on long-time experience in data management and in close cooperation with potential users. For this purpose, MOSAIC seeks to cooperate with newly initiated and already existing cohort studies and registries to acquire knowledge about specific needs and perceptions resulting from individual settings. The first set of tools, already available from the MOSAIC portal, is derived from solutions developed within existing projects [6–9]. Taking usability improvements for individual software tools as one example, the target was to reduce the number of configuration steps and to simplify the integration into existing infrastructures. Ease-of-use is improved by deploying web-based graphical user interfaces and by adding documentation (e.g. quick start guides, developer documentation and brochures).

To feature a high-level overview necessary for planning, designing and implementing CDM, a **web-based reference platform** [13] is offered. Depicting the necessary steps and typical issues occurring during the individual phases, the interactive reference platform aims to concentrate existing topical knowledge from the research community. For this purpose, based

**Table 1** Functional and non-functional elements of CDM

	Core Element	Description
Functional requirements	Acquisition	Specification of data sources, data formats, interfaces and data transfer methods
	Processing and integration	Implementation of data integration processes, metadata enrichment and data quality assurance
	Standardized storage	Planning and implementation of a generic data model and provisioning of the necessary IT infrastructure
	Use and access	Options for data exploration, coordination of the data request process, and the import/export of data
Non-functional requirements	Integrity	Comprises measures for guaranteeing data consistency, security and protection
	Transparency	Surveillance plus continuous monitoring of processes and measures ensuring end-to-end traceability and reproducibility
	Extensibility	Hardware/Software must be readily extensible to accommodate expansion of the study and a higher level of requirements
	Availability	High level of long-term reliability for collection, storage, archiving and provisioning of data
	Data protection	Separation of identifying and medical data at the earliest possible juncture in accordance with data protection legislation. If participants must be uniquely identifiable, the informed consent and revocation documentation must be effectively managed. Pseudonymisation and anonymisation of the data must be possible.
	IT and information security	As defined by the BSI’s Baseline Protection Catalogue [11]. Comprises measures for authentication, authorisation, secure data transmission, encrypted data storage and the hardening of network infrastructure against unauthorized access.

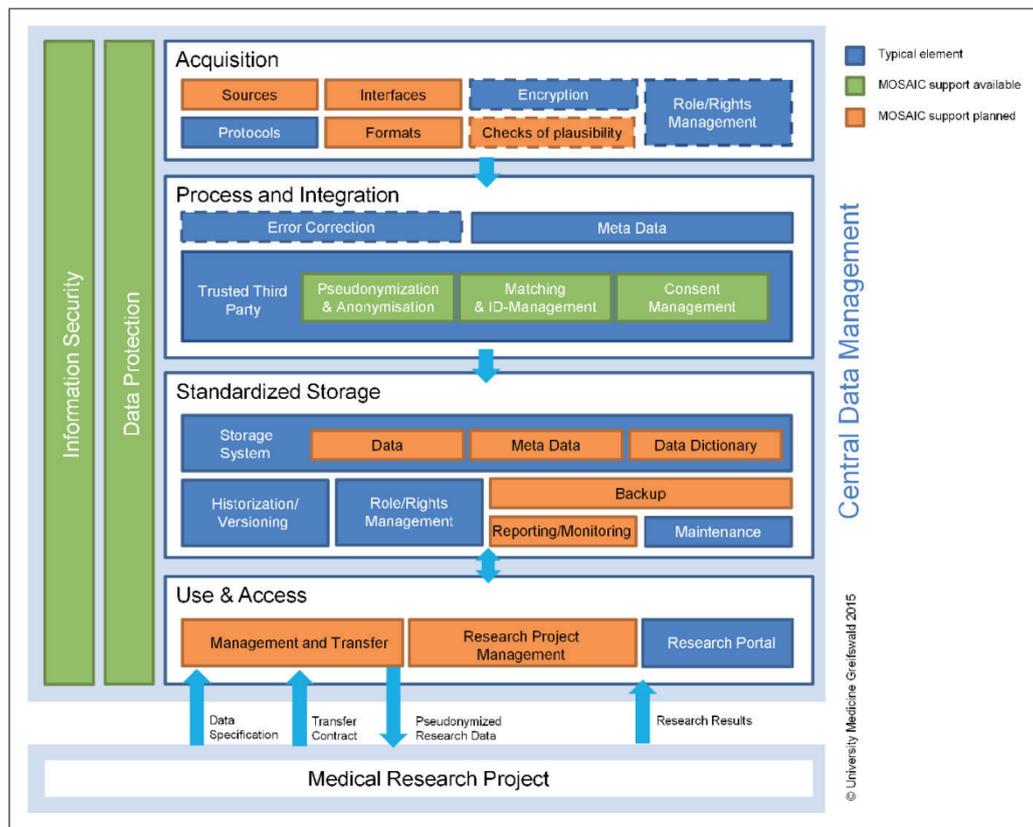


Figure 3 Functional requirements for CDM in an epidemiological context: typical responsibilities and specialized MOSAIC tools covering selected issues

on structured questions and answers, users are encouraged to rate existing solutions and to share their issues and experiences. The reference platform is based on the open source question-and-answer framework Question2Answer [14] and is completely integrated into the MOSAIC web-portal.

Setting up a new epidemiological project typically includes drafting a **data protection concept** note. To provide a starting point and to prevent missing essential data protection issues, MOSAIC shares a **document template** (DOCX-Format) to simplify the process of writing. Structured questions and recommendations guide potential authors towards the integration of study-specific attributes into a pre-defined,

pre-formatted concept paper. Alongside a conventional and directly deployable version of the document, an additional “generator” version enables interactive modification of the document template structure regarding mandatory, selectable and optional components beforehand. The document-generator utilizes several preformatted Microsoft Word building blocks, macros and configuration scripts optimized for the use with Microsoft Office 2010.

To avoid duplicate participant entries, the **ID Management solution E-PIX** (Enterprise Identifier Cross Referencing) applies the Fellegi-Sunter algorithm [15] and the Levenshtein distance. The independent software module allows for unambiguous participant management and effi-

cient aggregation of medical research data from federated study centres. Additionally, the correction of potential synonym errors is supported (i.e. false-negative record linkage). The E-PIX, as well as the subsequent tools, applies a service-oriented architecture to provide all functionalities via web services. It was developed using Java EE and several development frameworks, including PrimeFaces [16] for a web-based user-interface.

Before research data can be collected or provided within a cohort study or registry, legal conditions require checking for available consents or revocations from the specific participant. To manage both digital and paper-based informed consents MOSAIC offers the **Consent Management**

**solution gICS** (generic Informed Consent Administration Service) [17], which allows to check for various policies and modules of a consent automatically in real time. Comprehensive drafting of the consents and their validation through an ethics committee is assumed beforehand.

In cohort studies and registries, data storage requires pseudonymisation of each data record. Mostly the use of different pseudonyms in separate study centres or for different data categories (e.g. specimen, image data and medical data) is mandatory. The software module **gPAS** (generic Pseudonym Administration Service) [18] **generates and administers appropriate pseudonyms** using non-deterministic pseudonyms for arbitrary alphanumeric sequences. Additionally it allows defining domain-specific alphabets and generator algorithms as required and offers functions for de-pseudonymisation and anonymisation.

Depicting only selected disciplines in CDM for epidemiological research yet, the presented tools primarily facilitate ensuring conformity to legal data protection regulations. Current and future work within MOSAIC focusses on remaining aspects of CDM (►Figure 3, “MOSAIC support planned”). This comprises a solution for standardised storage, including an EAV-based metadata repository allowing for free definition of hierarchies between study items and the suitable research data repository. Supplementary ready-to-use examples will be provided to demonstrate how to use the storage solution with the open source EDC-Software OpenClinica [19], how to integrate research data from external devices (e.g. laboratory devices) and recommendations as well as checklists for an enhanced data protection strategy. Furthermore example reports for a basic quality assurance are elaborated using the open source statistical computing library R [20] allowing for an automated evaluation of metric and categorical study items. Also the preparation of document templates, in order to support the procedure of data provision, is intended.

All MOSAIC tools are made available on the project portal ([mosaic-greifswald.de](http://mosaic-greifswald.de)) and the mosaic subversion repository [21] under open source licensing. This applies to templates and documentation

(CC BY 4.0 [22]) as well as software (AGPLv3 [23]).

## 5. Discussion

The growing relevance of large-scale research networks in the scientific community makes the deployment of uniform methodologies and the generation of homogeneous data essential, since this allows for efficient data pooling and facilitates comparability. Consequently, work on the MOSAIC project complies with the recommendations of the German Research Foundation (DFG) for secure storage and provisioning of digital research data [2].

Previous work by Fraser et al. [24] has shown that the re-use of specially developed tools is capable of cutting costs and effort involved for data collection in separate studies. Moreover, utilizing the proposed set of MOSAIC tools increases the efficiency and standardisation of individual work within cohort studies and registries resulting from methodology streamlining. This standardisation facilitates an increased data quality [25].

The MOSAIC project's approach to provide modular solutions for specific requirements addresses the need for practical tools, in terms of readily deployable templates and software, in order to support planning, design and implementation of CDM for cohort studies and registries. Especially in all cases where the scientific environment offers only marginal experience in data management, lacks access to IT personnel or suffers from insufficient resources in terms of software development.

Implementation of the respective tools demands precise knowledge of developments in the research community in order to avoid conducting primary development work in parallel to existing solutions. In case proven solutions existed for well-defined issues, the extent to which these solutions meet the requirements of epidemiological cohort studies and registries and the degree of applicability was assessed.

The TMF already provides a Guideline for Data Protection in Medical Research Projects (published in 2006, updated in 2014) [12]. Though it presents an introduc-

tion, guidance and recommendations to all aspects of data protection and ethical issues, it lacks an easy to use document template, which actually supports the respective author to write a data protection concept note. This aspect is well addressed by the MOSAIC template (published 2013), providing a structured starting point for essential typical data protection issues to be answered in the latter process. Using a question-based approach supports the application in various study or registry scenarios. Applying pre-worded text-blocks would narrow the scope of potential users, e.g. the data protection template of the Open source Registry System for Rare Diseases in the EU (published 2014 within OSSE, [26]) fits only for registries using the OSSE Software, but actually accelerates drafting a data protection concept note.

Several tools for the management of participants exist in the scientific community. For example the TMF PID-Generator (published 2005) [27] allows for merging participant identifying data from federated study sites even if the data sets are incomplete or faulty. The Java-based Mainzel List (published 2013) [28] aims for a more contemporary approach. It comes with a likely set of functionality, but with an easy to use REST-interface, which facilitates simplified system integration. However, unlike the E-PIX (published 2014), both systems are not yet capable of managing multiple local identifiers and identities for each participant and lack a graphical user interface to support the respective user in detecting and solving possible synonym errors.

Also for pseudonymisation and de-pseudonymisation a well-established tool exists. The TMF Pseudonymisation Service (PSD, published 2010, [29]) generates pseudonyms with a fixed length and alphabet (depending on the selected encryption method) using a synchronous algorithm. Thus the storage of associated value-pairs (original value and associated pseudonym) is not necessary. As a consequence the anonymisation of a medical dataset by simply deleting the association, which connects original value and pseudonym, is not possible. However gPAS (published 2013) [18] provides this mechanism and additionally facilitates the creation of pseudonym hierarchies. The generation of multiple pseu-

onyms for one participant allows a context-related pseudonymisation e.g. to use specific pseudonyms for different data sources (eCRF, specimen, MRI) or for data provision during the use and access process.

Unlike the TMF Informed Consent Wizard (published 2007) [30], gICS does not provide textual support for drafting a document. Neither gICS focusses clinical practise nor is it limited to a simplified file-based approach of HL7-documents like the Consent Management Suite (COMS, published 2011) [31]. gICS (published 2014) focusses on the management of informed consent documents using re-usable policies to build modules, providing the necessary service functionalities to support automatic checks within seconds, e.g. whether a participant allows to share his specimen or has revoked his consent.

The developed MOSAIC tools will be evaluated in cooperation with the users in order to assess their efficacy and acceptance. One particular evaluation element is a dynamic review from a functional perspective. This review is conducted by deploying the tools to current and future external and internal projects enabling direct tool modifications and enhancements. Another element is tool testing conducted by external partners. The template for drafting data protection concepts has already been used in several research projects such as the TORCH registry of the DZHK or the MonDAFIS Study of the Centre for Stroke Research Berlin at the Charité (CSB). The modules for ID-Management, pseudonym administration and informed consent administration have recently been used to set up trusted third parties as a substantial part of CDM in compound projects such as the German Centre for Cardiovascular Research [9] or the German National Cohort [8]. In addition the MonDAFIS Study of the CSB uses the ID-Management solution E-PIX altogether with the open source EDC-Solution REDCap [32].

## 6. Conclusions

Applying common phases of a cohort study, this paper has deduced core elements of CDM for cohort studies and

registries in terms of functional and non-functional requirements. Associated challenges were highlighted, including requirements for IT security and legal data protection regulations.

In summary, CDM comprises the implementation of extensive technical and organisational means for the collection, processing, storage and provisioning of medical research data in accordance with the individual requirements specified by an epidemiological cohort study. Early incorporation of CDM features in study planning permits successful implementation before data collection commences. Furthermore CDM facilitates a process of systematic improvements in terms of data quality and availability while compliance with statutory legal frameworks is ensured.

This paper has introduced the MOSAIC project as a potential resource to support the implementation of CDM in epidemiological research with a first set of open source tools. Initial responses from the scientific community document how the concept of tool re-usability can be implemented while preserving associated benefits (cf. [24]). An important lesson learned is that the simplest solutions (e.g. a document template) awake great public interest.

Nonetheless, the MOSAIC tools will be unable to cover each requirement for CDM from an epidemiological perspective. Initially, no support for specimen management or for the linkage, import and processing of secondary data will be offered. In addition, no provision of servers or facilities for long-term data preservation is intended at present. MOSAIC also does not act as the commissioning instance for the services of a Trusted Third Party. On the other hand, technical support requests related to the tools as provided can be submitted via the project portal at any time.

For the remainder of its term, the MOSAIC project will focus on developing additional tools, which will continuously be made available to the research community via the MOSAIC project portal ([mosaic-greifswald.de](http://mosaic-greifswald.de)).

## Acknowledgments

This research is funded by the German Research Foundation (DFG) as a part of the

research grant programme “Information infrastructure for research data” (grant number HO 1937/2-1)

## References

- Higgins S. The DCC Curation Lifecycle Model. *The International Journal of Digital Curation* 2008; 3 (1): 134–140. doi: 10.2218/ijdc.v3i1.48.
- Committee on Scientific Library Services: Subcommittee on Information Management. Recommendations for Secure Storage and Availability of Digital Primary Research Data (Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten). [Online]. Bonn 2009 [cited 2015 02 10]. Available from: [http://www.dfg.de/download/pdf/foerderung/programme/lis/ua\\_inf\\_empfehlungen\\_200901\\_en.pdf](http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901_en.pdf).
- Jensen U. Guidelines for the management of research data (Leitlinien zum Management von Forschungsdaten). Technical Report. Köln: GESIS – Leibniz Institute for Social Sciences, Social Sciences; 2012.
- Winter A, Funkat G, Haeber A, Mauz-Koerholz C, Pommerening K, Smers S, et al. Integrated Information Systems for Translational Medicine. *Methods Inf Med* 2007; 46 (5): 601–607. doi: 10.1160/ME9063.
- Sariyar M, Borg A, Pommerening K. Evaluation of Record Linkage Methods for Iterative Insertions. *Methods Inf Med* 2009; 48 (5): 429–437. doi: 10.3414/ME9238.
- Völzke V, Alte D, Schmidt CO, Radke D, Lorbeer R. Cohort Profile: The Study of Health in Pomerania. *International Journal of Epidemiology* 2011; 40 (2): 294–307. doi: 10.1093/ije/dyp394.
- Grabe H, Assel H, Bahls T, Dörr M, Endlich K, Endlich N, et al. Cohort profile: Greifswald approach to individualized medicine (GANI\_MED). *Journal of Translational Medicine* 2014; 12 (144). doi: 10.1186/1479-5876-12-144.
- The German National Cohort (Nationale Kohorte e.V.). The German National Cohort Website. [Online]. 2014 [cited 2015 02 10]. Available from: [http://www.nationale-kohorte.de/content/Datenschutzkonzept\\_130314.pdf](http://www.nationale-kohorte.de/content/Datenschutzkonzept_130314.pdf).
- German Centre for Cardiovascular Research (DZHK). *dzhk.de*. [Online]. 2015 [cited 2015 02 10]. Available from: <http://dzhk.de/>.
- Meyer J, Ostrzinski S, Fredrich D, Havemann C, Krafczyk J, Hoffmann W. Efficient data management in a large-scale epidemiology research project. *Comput Methods Programs Biomed* 2012; 107 (3): 425–435. doi: 10.1016/j.cmpb.2010.12.016.
- German Federal Office for Information Security (Bundesamt für Sicherheit in der Informationstechnik). BSI-Standard 100-2 IT-Grundschutz-Vorgehensweise. Bonn: German Federal Office for Information Security (Bundesamt für Sicherheit in der Informationstechnik); 2008.
- Pommerening K, Drepper J, Helbing K, Ganslandt T. Guideline for Data Protection in Medical Research Projects: TMF's generic solutions 2.0. 1st ed. Berlin; 2014.

13. MOSAIC. The Reference Portal plan.Tau. [Online]. 2014 [cited 2015 02 24]. Available from: <https://mosaic-greifswald.de/qa/info>.
14. Greenspan G. Question2Answer. [Online]. [cited 2015 2 24]. Available from: <http://www.question2answer.org/>.
15. Fellegi I, Sunter A. A theory for record linkage. *Journal of the American Statistical Association* 1969; 64 (328): 1183–1210. doi: 10.1080/01621459.1969.10501049.
16. PrimeTek. PrimeFaces. [Online]. 2014 [cited 2015 2 24]. Available from: <http://primefaces.org/>.
17. Bahls T, Liedtke W, Geidel L, Langanke M. Ethics Meets IT: Aspects and elements of Computer-based informed consent processing. In Fischer T, Langanke M, Marschall P, et al., editors. *Individualized medicine, ethical, economical and historical perspectives*. Springer; 2015. pp 209–229.
18. Geidel L, Bahls T, Hoffmann W. Ein generisches Pseudonymisierungswerkzeug als Modul des Zentralen Datenmanagements medizinischer Forschungsdaten. In: Löffler M, Riedel-Heller S, editors. *Abstractband 8th Annual Conference of the German Society for Epidemiology (DGEpi) e.V. and 1st International LIFE Symposium (Abstractband 8. Jahrestagung der Deutschen Gesellschaft für Epidemiologie und 1. Internationales LIFE Symposium)*. Leipzig; 2013. pP 245–246.
19. OpenClinica, LLC. Open Clinica – Open Source for Clinical Research. [Online]. 2015 [cited 2015 2 24]. Available from: <https://community.openclinica.com/>.
20. The R Foundation. The R Project for Statistical Computing. [Online]. 2015 [cited 2015 2 24]. Available from: <http://www.r-project.org/>.
21. The MOSAIC Project. MOSAIC Subversion Repository. [Online]. 2015 [cited 2015 2 24]. Available from: <http://www.mosaic-greifswald.de/submin>.
22. Creative Commons. Creative Commons Attribution 4.0 International License. [Online]. 2013 [cited 2015 02 10]. Available from: <http://creativecommons.org/licenses/by/4.0/>.
23. Free Software Foundation. GNU Affero General Public License. [Online]. 2007 [cited 2015 02 10]. Available from: <http://www.gnu.org/licenses/agpl-3.0.html>.
24. Fraser H, Thomas D, Tomaylla J, Garcia N, Lecca L, Murray M, et al. Adaptation of a web-based, open source electronic medical record system platform to support a large study of tuberculosis epidemiology. *BMC Medical Informatics and Decision Making* 2012; 12 (125). doi: 10.1186/1472-6947-12-125.
25. Sariyar M, Borg A, Heidinger O, Pommerening K. A practical framework for data management processes and their evaluation in population based medical registries. *Informatics for Health and Social care* 2013; 38 (2): 104–119. doi: 10.3109/17538157.2012.735731.
26. Muscholl M, Lablans M, Wagner T, Ückert F. OSSE – open source registry software solution. *Orphanet Journal of Rare Diseases* 2014; 9 (Suppl 1). doi: 10.1186/1750-1172-9-S1-O9.
27. TMF e.V. The TMF PID-Generator. [Online]. 2014 [cited 2015 02 10]. Available from: [http://www.tmf-ev.de/Themen/Projekte/V015\\_01\\_PID\\_Generator.aspx](http://www.tmf-ev.de/Themen/Projekte/V015_01_PID_Generator.aspx).
28. Muscholl M, Lablans M, Borg A, Ückert F. Integration des Identitätsmanagements für Forschungsdatenbanken in ETL-Prozesse am Beispiel der Mainzer Patientenliste. In: *GMDS 2013. 58. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (GMDS)*; 2013; Lübeck. doi: 10.3205/13gmds052.
29. TMF e.V. The TMF pseudonymization service (PSD), a tool for the reversible pseudonymization of medical research data. [Online]. 2014 [cited 2015 02 16]. Available from: [http://www.tmf-ev.de/Projekte/TMFProjekte/V000\\_01\\_PSD.aspx](http://www.tmf-ev.de/Projekte/TMFProjekte/V000_01_PSD.aspx).
30. TMF e.V. Informed Consent – Software wizard of the TMF provides practical support. [Online]. 2007 [cited 2015 01 29]. Available from: <http://www.tmf-ev.de/News/articleType/ArticleView/articleId/223.aspx>.
31. Heinze O, Birkle M, Köster L, Bergh B. Architecture of a consent management suite and integration into IHE-based regional health information networks. *BMC Medical Informatics and Decision Making* 2011; 11 (58). doi: 10.1186/1472-6947-11-58.
32. Harris P, Taylor R, Thielke R, Payne J, Gonzalez N, Conde J. Research electronic data capture (RED-Cap) – A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics* 2009; 42 (2): 377–381. doi: 10.1016/j.jbi.2008.08.010.

## METHODOLOGY

## Open Access



# A workflow-driven approach to integrate generic software modules in a Trusted Third Party

Martin Bialke<sup>1\*</sup>, Peter Penndorf<sup>1,3</sup>, Tim Wegner<sup>2</sup>, Thomas Bahls<sup>1,3</sup>, Christoph Havemann<sup>1</sup>, Jens Piegsa<sup>1</sup> and Wolfgang Hoffmann<sup>1,3</sup>

## Abstract

**Background:** Cohort studies and registries rely on massive amounts of personal medical data. Therefore, data protection and information security as well as ethical aspects gain in importance and need to be considered as early as possible during the establishment of a study. Resulting legal and ethical obligations require a precise implementation of appropriate technical and organisational measures for a Trusted Third Party.

**Methods:** This paper defines and organises a consistent workflow-management to realize a Trusted Third Party. In particular, it focusses the technical implementation of a Trusted Third Party Dispatcher to provide basic functionalities (including identity management, pseudonym administration and informed consent management) and measures required to meet study specific conditions of cohort studies and registries. Thereby several independent open source software modules developed and provided by the MOSAIC project are used. This technical concept offers the necessary flexibility and extensibility to address legal and ethical requirements of individual scenarios.

**Results:** The developed concept for a Trusted Third Party Dispatcher allows mapping single process steps as well as individual requirements and characteristics of particular studies to workflows, which in turn can be combined to model complex Trusted Third Party processes. The uniformity of this approach permits unrestricted re-combination of the available functionalities (depending on the applied software modules) for various research projects.

**Conclusion:** The proposed approach for the technical implementation of an independent Trusted Third Party reduces the effort for scenario specific implementations as well as for maintenance. The applicability and the efficacy of the concept for a workflow-driven Trusted Third Party could be confirmed during the establishment of several nationwide studies (e.g. German Centre for Cardiovascular Research and the National Cohort).

**Keywords:** Medical data management, Data protection, Informed consent, Pseudonyms, Record linkage

## Background

Epidemiological research in the context of cohort studies and registries becomes increasingly cooperative and often requires multi-site acquisition of extensive medical data. As a consequence research becomes more and more networked regarding communication, information

exchange and cross-coordination between participating research institutions, laboratories and imaging facilities.

For these reasons, legal aspects of data security and information protection significantly gain in importance. This concerns the written informed consent of potential participants, which is mandatory for acquiring medical data for research purposes from an ethical point of view. On the national level legal principles like data avoidance and frugality [§3a of the German Federal Data Protection Act (Bundesdatenschutzgesetz, BDSG)] as well as requirements for the separation of identifying data from further personal data (§40 BDSG) need to be accounted

\*Correspondence: martin.bialke@uni-greifswald.de

<sup>1</sup>Institute for Community Medicine, Section Epidemiology of Health Care and Community Health, University Medicine Greifswald, Ellernholzstr. 1-2, 17487 Greifswald, Germany

Full list of author information is available at the end of the article



© 2015 Bialke et al. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

for. International legislation includes the “Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data” (Council of Europe) [1], the “EU legal framework on the protection of personal data” (European Commission) [2] and the “Declaration of Helsinki” (World Medical Association) [3]. The resulting legal and ethical obligations require effective solutions realizing all necessary measures for data protection and IT security.

In Germany the *Technology, Methods and Infrastructure for Networked Medical Research* (TMF) provided a guideline [4] proposing a Trusted Third Party (TTP) to address typical challenges in data protection and ethics. Following the TMF-specification a TTP requires an informational separation of powers by separating person identifying information (PII) and medical information from a technical as well as from an organizational perspective. This includes an electronic identity-management and should be supplemented by a secure pseudonymisation mechanism [4]. Following this definition, a TTP is described as a combination of technical as well as organisational measures and shall comply with fundamental principles according to data protection rules for IT-solutions [4]. Moreover, the guideline demands the TTP to be legally, staff-wise and spatially autonomous and independent.

It is of importance that the employees of a TTP (the data trustee) do not depend on the institutions which are providing or processing the research data. In particular the employees need to be independent in terms of their contracts, incomes, duties, work hours and other operational aspects from all scientists of the project that they support. This can be realized either in a separate legal organisation or on a contract level. According to TMF guidelines [4] and legal reports [5] as well as the Federal Data Protection Act (cf. §28 BDSG) the processing of data on a contract level prevents a sufficient informational separation of powers. From an organisational perspective the TTP requires a functional transfer (transfer of full responsibilities for data processing) in order to be independent of instructions from the initiators of a research.

Along these guidelines an independent TTP was established at the Institute for Community Medicine at the University Medicine Greifswald to exclusively handle participant identifying data, which have been separated from all further informative variables including meta-data. The TTP centrally provides the necessary data protection functionalities and measures for various studies and registries. Figure 1 presents an overview of a typical TTP infrastructure and involved stakeholders in the context of research data management.

This paper focusses the technical implementation of a TTP. The goal is to define and organise a consistent

workflow-management within the TTP. Allowing to increase reusability for individual TTP scenarios, the workflow approach shall reduce the effort for implementation and maintenance.

## Methods

### Assembling a modular Trusted Third Party

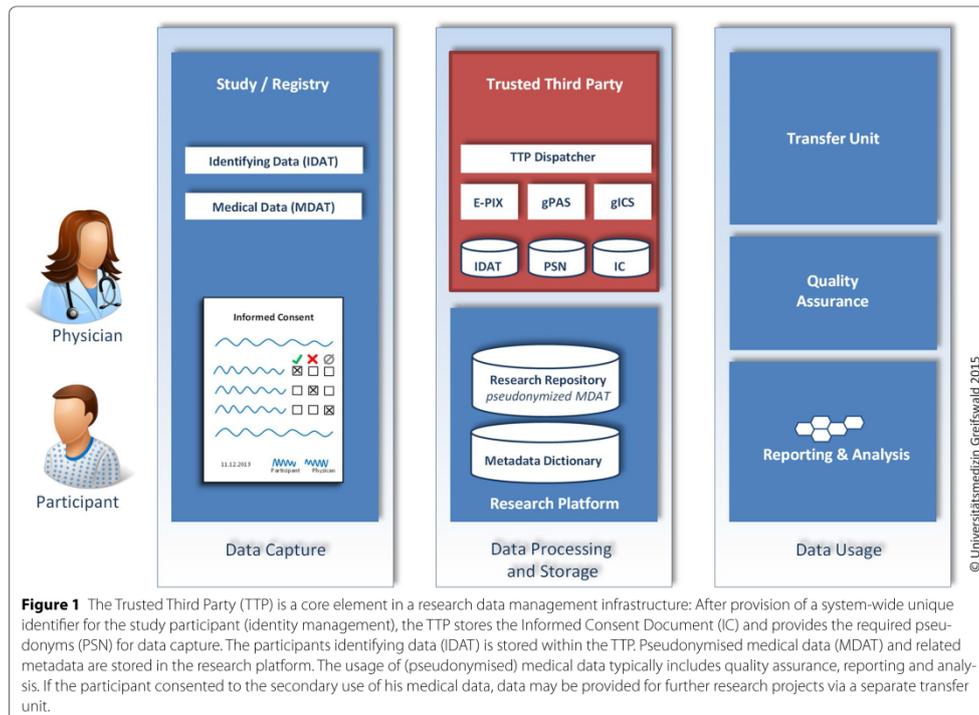
The spectrum of tasks of a data trustee includes the management of identities, informed consents and the generation of pseudonyms. Additionally, the data trustee supports the matching of personal data from population registries and further external data sources.

An *identity management* is required to manage participants and assigned participant identities. It includes probabilistic matching algorithms for an efficient and fault-tolerant record-linkage. Furthermore, it comprehends the provision and management of appropriate pseudonyms for each set of identities. Especially in prospective cohort studies and registries compiling variations in the identifying data of a participant (a so called identity), e.g. different spelling in a participant's name, need to be stored.

To ensure compliance to the principles of informational self-determination [4], the participant has to be able to consent to several aspects of data processing. Within the TTP the *management of informed consents* includes the provision of patient information documents, the consent itself and a monitoring of various types of revocations. For digital processing informed consent documents are depicted as modular examinable policies and modules and are combined with additional data like electronic signatures, dates and organisational information. This modular informed consent allows for verifiable as well as contemporary statements, whether for example the processing of a participant's data, the secondary use of collected data or the specimen-collection is legitimate or not.

The efficient generation and *administration of pseudonyms* within the TTP is a key functionality when medical scientific data needs to be processed and permanently stored. In order to provide scientific data for research projects and secondary use, the data has to be pseudonymised secondarily or be anonymised. In some cases an anonymisation is not applicable. Follow-up investigations, the communication of incidental findings or the linkage of secondary data require the pseudonymisation to be reversible in order to retrieve the corresponding participants for further contact.

For the implementation of the independent TTP several open source software modules are used. Following the basic concepts and processes described by the TMF [4], the MOSAIC project [6] (funded by the German Research Foundation (HO 1937/2-1)) has developed a



set of practical tools to address data protection challenges and to provide support for the implementation of a data management in epidemiologic research projects. These free software tools (E-PIX, gICS, gPAS) facilitate the principles of “privacy by design” [7] and use uniform technical standards. Moreover these tools provide a service-oriented architecture and consistent graphical user interfaces.

The *E-PIX* (*Enterprise Patient Identifier Cross Referencing*) [8] allows a precise identity management and supports the data trustee to distinguish participants sustainably based on their identifying data (IDAT). It follows the principles of a Master Person Index. This ensures a participant to exist only once in the linkage database based on demographic information [9]. The completely service-based software module generates a unique identifier for every managed participant and allows solving ambiguous matching cases interactively using a web-based graphical interface. The equally modular solution *gPAS* (*generic Pseudonym Administration Service*) [10] adopts similar technical approaches and provides domain-specific pseudonym creation,

de-pseudonymisation and anonymisation functionalities. The utilisation of *gICS* (*generic Informed Consent Service*) [11] completes the set of TTP tools. It facilitates the management of digital informed consent documents and allows automatable checks for consent validity and revocations [11]. Modular informed consents are defined, based on examinable policies and re-usable modules.

The simultaneous use of the MOSAIC software modules E-PIX, gICS and gPAS allows implementing basic requirements of an independent TTP. The administration of participant identities, informed consents and pseudonyms can be performed using graphical web interfaces. However, due to their modular design there is no direct communication among these components. In order to realize more complex workflows, a manual intervention of the data trustee is necessary in many tasks. For example, if a new participant is recruited, it is necessary to assign a unique identifier based on his IDAT (identity management), to pseudonymise this unique identifier (pseudonym administration) and to return the generated pseudonym in order to start capturing the medical data within the study site.

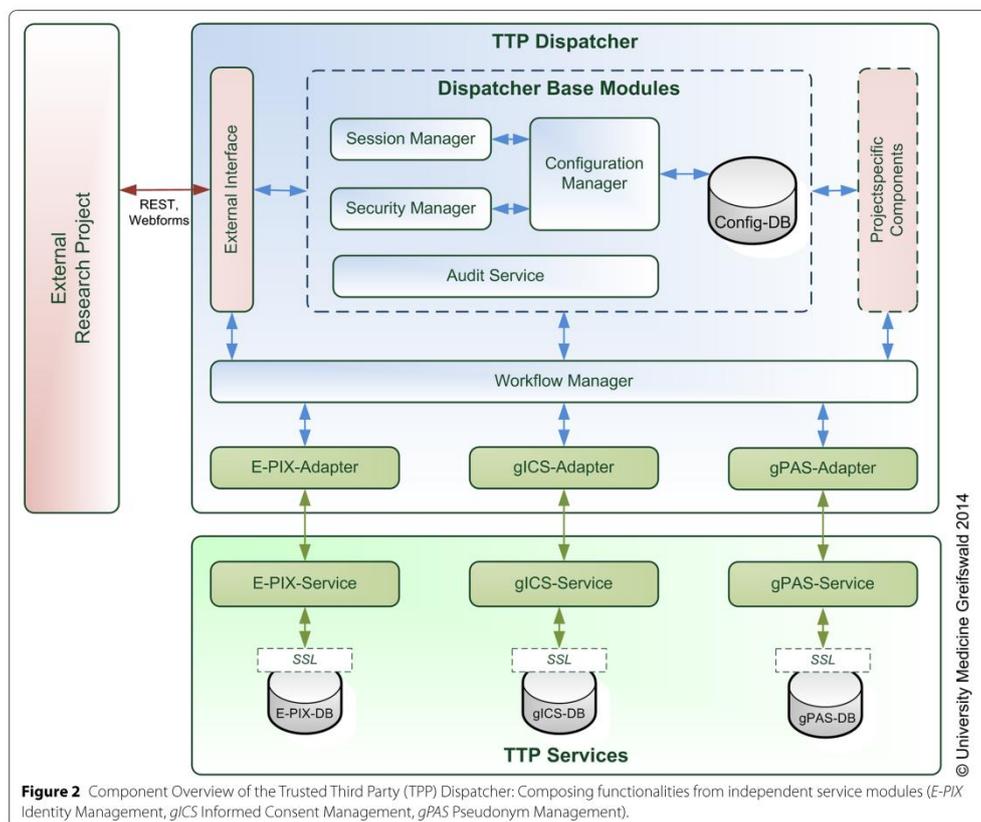
Most widely automating the communication between the software-modules E-PIX, gPAS and gICS through well-defined workflows reduces the number of necessary manual interventions of the data trustee. Only a small number of crucial decisions remains, where a human interaction cannot be replaced (e.g. to evaluate and resolve possible matches).

**Extending flexibility to support individual scenarios**

The required communication between the previously described TTP services depends on the workflow of a specific cohort or registry and, hence, individual characteristics may differ from the typical scenario. In order to flexibly orchestrate the particular TTP services and to coordinate the corresponding communication between the services, a dispatcher has been developed. The *TTP Dispatcher* represents the conceptual continuation of a

request dispatcher, which was introduced in the GANI\_MED project. [12].

Figure 2 presents an overview of the implemented components and defines the specific data flows. The TTP Dispatcher consists of several modules and communicates with the previously described *TTP service* modules gPAS, gICS and E-PIX via corresponding *Service Adapters*. The service adapter connects the TTP Dispatcher with the respective module. This approach enhances the interoperability of the TTP implementation and reduces the technical effort to add further software modules if necessary. Moreover, using service adapters facilitates and standardises the dispatcher-internal handling of the utilized service functionalities. The *Configuration Manager* administrates project-specific and dispatcher-specific configurations in a separate database. This includes workflows, roles and rights as well as a set of individual



**Figure 2** Component Overview of the Trusted Third Party (TTP) Dispatcher: Composing functionalities from independent service modules (E-PIX Identity Management, gICS Informed Consent Management, gPAS Pseudonym Management).

parameters (e.g. for informed consents or data entry validation). The *External Interface* allows connections to selected functionalities for external use in order to integrate the TTP Dispatcher in external applications. It can be accessed using the representational state transfer protocol (REST) or web-forms. The external interface and concept is oriented towards an existing identity management solution (the "Mainzliste" [13]). The specification was extended in cooperation with the authors. Another key component is the *Session Manager*, which handles external requests and administrates all required information. In order to grant access to available functionalities for registered users and systems only, the *Security Manager* provides a basic role-and-rights-management. The TTP Dispatcher is not limited to the components in Figure 2. *Project specific components*, e.g. to process data from smart cards, sign pads for electronic signatures, individual web forms as well as additional databases, can be readily integrated. The administration of project specific workflows is handled by the *Workflow Manager*.

#### Flexibility through workflows

In terms of a TTP, a workflow technically describes a sequence of (parallel) processes and operations, starting with an input and ending with a defined outcome. Workflows are being used to control and process the necessary calls to the connected software modules E-PIX, gPAS and gICS. They are distinguished into groups. Basic workflows represent common tasks of a data trustee and are of relevance in most project scenarios, e.g. checking if a participant already exists in the management system or generating pseudonyms for a list of participant identifiers. Project-specific workflows describe all necessary individual processes and operations beyond, for example all required steps to automatically generate a pseudonym when a new participant is created in a study site based on his IDAT and a valid informed consent. The separation of basic and project-specific workflows allows a consistent approach for several implementations of the TTP Dispatcher. This architecture hereby supports portability to other research projects, reduces maintenances and improves the sustainability of a TTP implementation.

The technical description of each workflow is performed using Apache Camel [14]. Based on Enterprise Integration Patterns [15] routes can be defined using a domain specific language. Each route comes with at least two end-points (source and target, e.g. a simple file, a web-service or an internal process), which are expecting an input (e.g. objects, messages) and returning a result. These end-points are linked using a message channel and basic elements of the Apache Camel syntax.

The starting point of a workflow defines the origin of an object or message ("from"). Combined with several processing steps (manual input, conversion), simple criteria-based conditions and a target point ("to") a workflow is specified. The necessary information (e.g. unique identifiers, pseudonyms or informed consent data) is passed directly between the individual workflow steps. The TTP Dispatcher comes with a basic set of predefined workflows (see Table 1 for details), which can also provide a basis for additional project-specific workflows.

#### Results

An essential part for the technical establishment of a Trusted Third Party is the implementation of required dispatcher functionalities. In the past the necessary individual implementations for a cohort study or registry required up to 6 month of work.

Using the proposed workflow-driven approach allows to accomplish the necessary customisations within weeks, by reusing various basic workflows and combining them as intended. For example creating a new participant (cf. Figure 3) in a study site of the DZHK, based on his IDAT and a valid informed consent (A), requires the combination of three basic workflows for full process automation: For record linkage the given IDAT of the participant are passed to the corresponding basic workflow using the E-PIX service (B). The result is a unique identifier which will be pseudonymised using the gPAS-Service subsequently (C). Finally the pseudonymised informed consent document is stored using the gICS-Service (D) and the TTP Dispatcher returns the pseudonym to the study site. Within the study site, the capture of medical data can begin.

In case of an error, the workflow processing is interrupted. Among other information, the error message and the error origin are documented and returned to the respective study site.

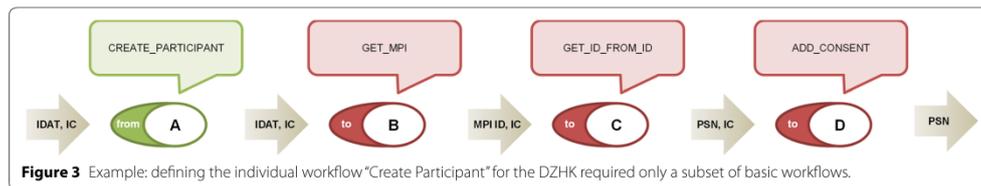
As the example demonstrates, a consistent workflow management allows easily linking available functionalities (E-PIX, gICS, gPAS) by reusing and combining predefined workflows. Thus the individual character of cohort studies and registries can be depicted straightforward. Moreover, study specific processes are most widely automatable and manual intervention of the data trustee could be essentially reduced.

#### Discussion

The components and workflows of a TTP vary according to their specific context. For example, the Central Clinical Cancer Registry in Mecklenburg-Western Pomerania [16] does not require a consent management and the German National Cohort [17] uses a specific

**Table 1 Overview of basic Trusted Third Party workflows**

Workflow	Description
get_mpi	Generate MPI ID for given IDAT using E-PIX-service
check_patient_exists	Check if a participant with given IDAT already exists in the E-PIX-database
get_id_from_id	Get a pseudonym for a given identifier (e.g. MPI ID) and vice versa using the gPAS-service
add_consent	Add a new informed consent (based on a template containing several modules and policies) for the given identifier using the gICS-service
check_consent_exists	Check if an informed consent for the given identifier exists in the gICS-database
query_consent	Query a list of policies and their consented state for a given informed consent identifier using the gICS-service
add_scan	Add a document scan to a previously defined informed consent using the gICS-service
update_participant	Update a participants IDAT already existing in the E-PIX-database
get_participant_by_mpi_id	Retrieve a participants IDAT from the E-PIX database identified by its MPI ID
add_participant_get_psn	Sequential workflow combining get_mpi and get_id_from_id
get_participant_by_psn	Sequential workflow combining get_id_from_id and get_participant_by_mpi_id



pseudonymisation for different sites and data categories (e.g. MRT, bio samples, web-forms).

Implementing a TTP on a basis of tools without uniform interfaces and with varying technical standards, demands an advanced IT knowledge and requires massive implementation efforts. That is why using the MOSAIC modules [8, 10, 11] seems to be a suitable and practical approach. Furthermore, using a workflow-based dispatcher in order to coordinate the specific functionalities [12] allows a consistent tailoring and adoption to new projects. Table 2 lists resulting advantages and disadvantages of the developed TTP Dispatcher.

Following the described approach for the technical implementation of a Trusted Third Party supports

compliance with project-specific legal data protection requirements in cohort studies and registries. But in order to exhaustively fulfil security, data protection, ethical and legal requirements [4], additional measures are necessary. Among others, this includes the institution of a data trustee, several dedicated rules, access controls for non-employees, separated network infrastructures and full client-capability on a technical and organisational level [18], resulting in the separated storage of participant identifying data for each supported study and registry. Moreover, regular internal and external audits have to be engaged involving both the institutional and the federal data protection officers.

The aim of the described TTP Dispatcher approach has a significant difference to existing IT platforms

**Table 2 Advantages and disadvantages of the developed Trusted Third Party Dispatcher**

Advantages	Disadvantages
Support for automation reduces susceptibility to errors and accelerates internal TTP processes	Initial configuration of the TTP Dispatcher requires professional IT support to set up mandatory databases and the application server
Integrated audit-and-trail-mechanisms ensure traceability and transparency of participating systems	Changes and updates in TTP Dispatcher core functions involve a determined update management including tests in the respective project, study or registry
Modular and adaptable workflows improve portability and re-usability	
Multi-client capability to manage large multi-site projects	
Interoperability of service components	

supporting clinical research, such as EHR4CR [19]. The proposed TTP approach focusses exclusively on the management of participant identifying data and related technical and organisational measures. Medical data is not processed within the TTP. EHR4CR focusses a widespread support to all steps of a clinical trial process instead. This includes the provision of information about new and running trials, several tools for data managers and investigators, the provision of query engines, recruitment software, an identity and access management, as well as a security framework. Unlike the proposed TTP approach, the EHR4CR IT platform stores aggregated medical information and patient identifying data does not leave the clinical context.

The concept of the TTP Dispatcher has already seen successful implementation in the German Centre for Cardiovascular Research (DZHK) [20] and the German National Cohort [17]. The resulting pattern can flexibly be adopted and easily be extended for reuse in future cohort studies and registries. The established TTP solutions are compatible to legal requirements of the "Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data" (Council of Europe) [1], the "Declaration of Helsinki" (World Medical Association) [3], the "EU legal framework on the protection of personal data" (European Commission) [2] as well as the "Treaty of Lisbon" (European Union) [21] and they are aligned to the previously mentioned recommendations of the TMF data protection concepts for medical research [4].

### Conclusions

During the recruitment of participants for cohort studies and registries particularly the acquisition, processing and storage of personal health data necessitate both compliance with ethical standards and stringent policies for data protection. For Germany, the resulting requirements for data management are compiled comprehensively in the guideline provided by the TMF [4]. Conformity is usually achieved by the implementation of a Trusted Third Party (TTP). However, the individual TTP implementation for different studies is associated with considerable high technical efforts that can be prohibitive in smaller studies or in institutions without a professional IT-department.

This paper demonstrates how generic software modules developed and provided by the MOSAIC project [8, 10, 11] can be deployed in order to meet essential TTP requirements. The concept of a workflow-driven dispatcher is introduced combining these modules in structured workflows, allowing for a free combination of separate functionalities. Single process steps can be easily implemented by concatenating corresponding function

calls and mapping them to workflows. The combination of multiple workflows enables an efficient conception and implementation of highly complex working procedures. Simultaneously the necessary effort for customisation is reduced to a minimum.

The proposed approach for the technical implementation of a TTP facilitates the necessary flexibility, portability and reusability for application in cohort studies and registries. This is achieved by mapping the individual requirements and characteristics of a particular study to pre-defined workflows. Reusability additionally benefits from the encapsulation of module logic and a uniform interface for all modules avoiding study specific modifications of individual modules or functionalities. The generic software modules connected with the workflow approach presented in this paper can easily be adopted to accommodate national and international requirements in terms of informed consent, identity management, pseudonymisation, data linkage and data transfer.

However, specification of a uniform interface for essential functionalities and parameters accounting for established standards and methods within the scientific community must still be considered time-consuming and labour-intensive. Future work will focus on further facilitating the establishment of an independent TTP, including workflow visualization, a generic module configuration independent from the deployed services, a graphical configuration tool for the configuration of the dispatcher and an extended central role-and-rights-management.

### Abbreviations

DZHK: Deutsches Zentrum für Herz-Kreislauf-Forschung (German Centre for Cardiovascular Research); E-PIX: Enterprise Patient Identifier Cross-referencing; gICS: generic Informed Consent Administration Service; gPAS: generic Pseudonym Administration Service; IDAT: identifying data; MDAT: medical data; PSN: pseudonym; MPI ID: unique identifier following concepts of a Master Patient Index; TMF: Technology, Methods and Infrastructure for Networked Medical Research; TTP: Trusted Third Party

### Authors' contributions

PP and TB were involved in the conception and design process of the TTP Dispatcher. MB, PP, TB, TW, CH, JP and WH drafted the manuscript. PP, TB, CH and WH revised it critically. PP and TB were responsible for the set-up of the IT infrastructure. All authors read and approved the final manuscript.

### Author details

<sup>1</sup>Institute for Community Medicine, Section Epidemiology of Health Care and Community Health, University Medicine Greifswald, Ellernholzstr. 1-2, 17487 Greifswald, Germany. <sup>2</sup>Institute of Applied Microelectronics and Computer Engineering, University of Rostock, Rostock, Germany. <sup>3</sup>German Centre for Cardiovascular Research (DZHK), Greifswald, Germany.

### Acknowledgements

The MOSAIC-Project is funded by the German Research Foundation (DFG) as a part of the research grant programme "Information infrastructure for research data" (grant number HO 1937/2-1).

**Compliance with ethical guidelines****Competing interests**

The authors declare that they have no competing interests.

Received: 9 February 2015 Accepted: 25 May 2015

Published online: 04 June 2015

**References**

1. Council of Europe. Convention for the protection of individuals with regard to automatic processing of personal data. In ETS No.108 1981 Strasbourg
2. European Commission (2012) Official website of the European Commission. [http://ec.europa.eu/justice/data-protection/document/review2012/com\\_2012\\_11\\_de.pdf](http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_de.pdf). Accessed 02 Oct 2014
3. World Medical Association (WMA) (2008) WMA Declaration of Helsinki—Ethical Principles for Medical Research Involving Human Subjects. In 59th WMA General Assembly 2008 Korea
4. Pommerening K, Drepper J, Helbing K, Ganslandt T, Müller T, Speer R et al (2014) Generic data protection concepts for medical research networks 2.0 (Leitfaden zum Datenschutz in medizinischen Forschungsprojekten. Generische Lösungen der TMF 2.0). Berlin. TMF e.V. 2014
5. Dierks C (2008) Legal evaluation of an electronic data trustee ship of the TMF (Rechtsgutachten zur elektronischen Datentreuhänderschaft der TMF, TMF-Produkt P052011). Berlin. 2008
6. The MOSAIC-Project (2014) Mosaic-Project Website. <http://mosaic-greifswald.de>. Accessed 1 Jun 2014
7. Schaar P (2010) Privacy by design. *Identity Inf Soc* 3(2):267–274. doi:10.1007/s12394-010-0055-x
8. The MOSAIC-Project (2014) ID-Management with E-PIX. <https://mosaic-greifswald.de/werkzeuge-und-vorlagen/id-management-e-pix.html>. Accessed 19 Sep 2014
9. Lenson C (1998) Building a successful enterprise master patient index: a case study. *Top Health Inf Manag* 19(1):66–71
10. Geidel L, Bahls T, Hoffmann W (2013) A generic pseudonymization tool as a module of Central Data Management for medical research data (Ein generisches Pseudonymisierungswerkzeug als Modul des Zentralen Datenmanagements medizinischer Forschungsdaten). In: Löffler M, Riedel-Heller S (eds) Abstractband 8th Annual Conference of the German Society for Epidemiology (DGEpi) e.V. and 1st International LIFE Symposium (Abstractband 8. Jahrestagung der Deutschen Gesellschaft für Epidemiologie und 1. Internationales LIFE Symposium). Leipzig, pp 245–246
11. Bahls T, Liedtke W, Geidel L, Langanke M (2015) Ethics meets IT: aspects and elements of computer-based informed consent processing. In: Fischer T, Langanke M, Marschall P, Michl S (eds) *Individualized medicine: ethical, economical and historical perspectives*. Springer, Cham, pp 209–229. <http://www.springer.com/biomed/book/978-3-319-11718-8>
12. Grabe H, Assel H, Bahls T, Dörr M, Endlich K, Endlich N et al (2014) Cohort profile: Greifswald approach to individualized medicine (GANI\_MED). *J Transl Med* 12:144. doi:10.1186/1479-5876-12-144
13. Lablans M, Borg A, Ückert F (2015) A RESTful interface to pseudonymization services in modern web applications. *BMC Med Inform Decis Mak* 15:2. doi:10.1186/s12911-014-0123-5
14. The Apache Software Foundation (2014) Apache Camel Homepage. <http://camel.apache.org/>. Accessed 20 Oct 2014
15. The Apache Software Foundation (2014) Enterprise Integration Patterns. <http://camel.apache.org/enterprise-integration-patterns.html>. Accessed 20 Oct 2014
16. ZKKR-MV (2014) Central Clinical Cancer Registry in MV (Zentrales klinisches Krebsregister Mecklenburg-Vorpommern). <http://web1-zkk.rzkk.med.uni-greifswald.de/>. Accessed 29 Sep 2014
17. The National Cohort (Nationale Kohorte e.V.) (2015) <http://www.nationale-kohorte.de/content/9.2-DS-Konzept-Treuhandstelle-NAKO-VI-01-2015-01-11.pdf>. Accessed 20 May 2015
18. Conference of the Federal and State Data Protection Officers—Working group for technical and organisational data protection issues (2012) Guide to client-capability (Technische und organisatorische Anforderungen an die Trennung von automatisierten Verfahren bei der Benutzung einer gemeinsamen Infrastruktur—Orientierungshilfe Mandantenfähigkeit). <http://www.baden-wuerttemberg.datenschutz.de/wp-content/uploads/2013/04/Mandantenfähigkeit.pdf>. Accessed 24 Apr 2015
19. Doods J, Bache R, McGilchrist M, Daniel C, Dugas M, Fritz F (2014) Piloting the EHR4CR feasibility platform across Europe. *Methods Inf Med* 53(4):264–268. doi:10.3414/ME13-01-0134
20. German Centre for Cardiovascular Research (DZHK) (2014) <http://dzhk.de/>. Accessed 11 Mar 2014
21. Conference of the Representatives of the Governments of the Member States. Treaty of Lisbon Amending the Treaty on European Union and the Treaty Establishing the European Community. In *Official Journal of the European Union* (2007/C 306/01) 2007 Lisbon, pp 1–228

**Submit your next manuscript to BioMed Central and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)



## B Ergänzungsmaterial

Projektname	Rückmeldung am	# Personen	# Unbearbeitete mögliche Matches	# Zusammengeführte mögliche Matches	# Getrennte mögliche Matches
CSB-MonDAFIS-Studie	12.01.2015	1	0	0	0
	14.04.2015	50	0	0	0
	10.06.2015	100	0	0	0
	11.09.2015	880	2	0	0
	25.04.2016	1.400	0	1	7
CSB-Berliner Vorhofflimmern Register	25.04.2016	775	0	0	0
DZHK e.V.	08.01.2015	8	0	0	0
	14.04.2015	81	0	0	0
	09.06.2015	150	0	0	0
	11.09.2015	394	2	0	0
	24.11.2015	618	4	0	0
	31.03.2016	971	7	0	0
GANI_MED	12.01.2015	5.437	621	54	9
	14.04.2015	5.463	48	658	68
	11.06.2015	5.464	1	703	72
	11.09.2015	5.529	0	703	73
Summative Evaluation Kifög M-V	11.06.2015	12.267	0	824	225
	11.09.2015	12.267	0	824	225
	20.01.2016	12.267	0	824	225
	31.03.2016	11.226	456	25	59
Nationale Kohorte	08.01.2015	83.376	4	155	1.846
	14.04.2015	137.775	2	158	3.156
	09.06.2015	167.706	2	303	3.633
	11.09.2015	241.258	0	510	6.018
	24.11.2015	317.748	0	528	7.032
	01.04.2016	407.680	22	776	8.291
Zentrales klinisches Krebsregister MV	08.01.2015	155.498	0	270	4.049
	09.06.2015	155.498	0	270	4.049
	11.09.2015	155.498	0	270	4.049

Tabelle 8 Übersicht der von Januar 2015 bis Mai 2016 von den E-PIX-Anwenderprojekten übermittelten Kennzahlen

Projektname	Rückmeldung am	# Pseudonyme	# Domänen	# Anonyme
<b>DZHK e.V.</b>	08.01.2015	48	11	16
	14.04.2015	268	11	16
	09.06.2015	475	12	16
	11.09.2015	1.207	12	16
	24.11.2015	1.900	12	16
	31.03.2016	2.998	12	16
	<b>GANI_MED</b>	12.01.2015	277.864	14
	14.04.2015	297.730	15	4
	11.06.2015	340.109	20	4
	11.09.2015	473.009	24	4
<b>Nationale Kohorte</b>	08.01.2015	234.654	15	0
	14.04.2015	371.631	9	0
	09.06.2015	455.100	9	0
	11.09.2015	650.289	9	0
	24.11.2015	1.202.286	12	0
	01.04.2016	1.875.591	12	0
<b>Verbrennungsregister</b>	03.05.2016	436	16	0
<b>Zentrales klinisches Krebsregister MV</b>	08.01.2015	155.498	1	0
	09.06.2015	155.498	1	0
	11.09.2015	155.498	1	0

Tabelle 9 Übersicht der von Januar 2015 bis Mai 2016 von den gPAS-Anwenderprojekten übermittelten Kennzahlen

Projektname	Rückmeldung am	# Einwilligungen	# Module	# Module (versioniert)	# Policies	# Widerrufe	# Unterzeichnete Policies	# Templates
<b>DZHK e.V.</b>	08.01.2015	8	20	20	42	-	192	7
	14.04.2015	81	26	26	43	-	1.929	9
	09.06.2015	150	34	34	47	-	3.648	10
	11.09.2015	395	34	34	47	3	9.583	10
	24.11.2015	634	34	34	47	7	15.430	10
	31.03.2016	1.013	34	34	47	9	24.619	10
<b>GANI_MED</b>	12.01.2015	5.274	84	6	6	42	20.707	36
	14.04.2015	5.401	85	6	6	-	21.207	37
	11.06.2015	5.401	85	6	6	-	21.207	37
	11.09.2015	5.401	85	6	6	57	21.207	37
<b>Nationale Kohorte</b>	08.01.2015	10.150	82	0	55	-	292.291	3
	14.04.2015	18.123	88	77	55	-	585.236	37
	09.06.2015	23.800	88	77	70	-	766.616	39
	11.09.2015	33.612	88	77	70	65	1.076.895	39
	24.11.2015	44.794	88	77	70	65	1.428.282	39
	01.04.2016	62.532	113	79	72	65	1.842.912	52

**Tabelle 10 Übersicht der von Januar 2015 bis Mai 2016 von den gICS-Anwenderprojekten übermittelten Kennzahlen. Die Anzahl der Widerrufe wurde erst seit September 2015 erhoben.**

## C Publikationsliste

### 2016

#### Eingeladene Vorträge

M. Bialke, T. Bahls, L. Geidel. **Umsetzungskonzepte und Möglichkeiten der Pseudonymisierung.** *TMF Workshop "Anonymisierung und Pseudonymisierung in Patientenversorgung und Forschung"*; Veranstalter: TMF; 2016 Mai 23; Berlin.

### 2015

#### Artikel (Zeitschrift)

M. Bialke, T. Bahls, C. Havemann, J. Piegsa, K. Weitmann, T. Wegner, W. Hoffmann. **MOSAIC. A modular approach to data management in epidemiological studies.** *METHODS OF INFORMATION IN MEDICINE*. Bd. 54, Nr. 4, S. 364-371, 8 2015

M. Bialke, P. Penndorf, T. Wegner, T. Bahls, C. Havemann, J. Piegsa, W. Hoffmann. **A workflow-driven approach to integrate generic software modules in a Trusted Third Party.** *JOURNAL OF TRANSLATIONAL MEDICINE*. Bd. 13, Nr. 176, 6 2015.

#### Konferenzbeiträge

M. Bialke, J. Piegsa, W. Hoffmann. **MOSAIC: Praktische Hilfestellung durch Vorlagen, Leitfäden und Empfehlungen.** (Poster) *14. Deutscher Kongress für Versorgungsforschung (DKVF)*; 2015 Okt 07; Berlin.

M. Bialke, J. Piegsa, L. Geidel, R. Schuldt, P. Penndorf, D. Langner, R. Wolff, T. Schwaneberg, R. Gött, R. Walk, T. Bahls, W. Hoffmann. **MOSAIC: Kostenfreie Werkzeuge für die epidemiologische Forschung.** (Poster) *14. Deutscher Kongress für Versorgungsforschung (DKVF)*; 2015 Okt 07; Berlin.

M. Bialke, R. Schuldt. **MOSAIC: praxis-orientierte Unterstützung für Kohortenstudien und Register.** (Vortrag) *14. Deutscher Kongress für Versorgungsforschung (DKVF)*; 2015 Okt 08; Berlin.

## 2014

### Konferenzbeiträge

M. Bialke, T. Bahls, C. Havemann, J. Piegsa, W. Hoffmann. **plan.Tau - Ein interaktives Werkzeug zur Konzeption eines zentralen Datenmanagements für die epidemiologische Forschung.** (Poster) *9. Jahrestagung der Deutschen Gesellschaft für Epidemiologie (DGEpi)*; 2014 Sep 17; Ulm.

M. Bialke, D. Langner, T. Bahls, C. Havemann, J. Piegsa, W. Hoffmann. **“Who am I? And if so, how many?” – The E-PIX as innovative system to manage person identities.** (Poster) *2nd Research Data Management Workshop*; 2014 Nov 27; Köln.

### Eingeladene Vorträge

M. Bialke. **MOSAIC: Ein Musterdokument zur Erstellung von Datenschutzkonzepten.** *TMF AG Datenschutz*; Veranstalter: TMF; 2014 Sep 16; Berlin.

M. Bialke. **MOSAIC: Praxisorientierte Werkzeuge und Vorlagen für Kohortenstudien und Register.** *TMF AG IT-QM*; Veranstalter: TMF; 2014 Nov 20; Berlin.

## 2013

### Konferenzbeiträge

M. Bialke, T. Bahls, C. Havemann, J. Piegsa, W. Hoffmann. **Zentrales Datenmanagement als Methode zur Verbesserung der Nachnutzbarkeit medizinischer Forschungsdaten.** (Poster) *8. Jahrestagung der Deutschen Gesellschaft für Epidemiologie (DGEpi) und 1. Internationales LIFE-Symposium*; 2013 Sep 26; Leipzig.

### Eingeladene Vorträge

M. Bialke, T. Bahls, C. Havemann, J. Piegsa, W. Hoffmann. **Zentrales Datenmanagement als Methode zur Verbesserung der Nachnutzbarkeit medizinischer Forschungsdaten.** *TMF-Workshop "IT in epidemiologischen Forschungsprojekten"*; 2013 Sep 25; Leipzig.

## **D Danksagung**

Die Erstellung dieser Arbeit wäre ohne bedingungslosen familiären Rückhalt und kontinuierliche wissenschaftliche Förderung nicht möglich gewesen. Mein Dank gilt meiner Frau, die mich trotz der Vielzahl schlafloser Nächte konsequent motivieren konnte. Gleichzeitig möchte ich auch Herrn Prof. Dr. Wolfgang Hoffmann danken, der mich stets anspornte und mir half, in der Wissenschaft Fuß zu fassen.

Darüber hinaus gilt mein ganz besonderer Dank all jenen, durch deren Unterstützung die MOSAIC-Werkzeuge die notwendige Qualität erreichen konnten: Lars Geidel, Dirk Langner, Peter Penndorf, Ronny Schuldt, Arne Blumentritt, Robert Gött, Robert Wolff, Rene Walk, Thea Schwaneberg, Henriette Rau, Kerstin Weitmann, Thomas Bahls, Christoph Havemann und Jens Piegsa.

Ihr habt maßgeblich zum Projekterfolg beigetragen.