



Pseudonymisierung und Record Linkage

NFDI ELSA Workshop 12.10.2022

Tobias Kussel

- ¹ Complex Data Processing in Medical Informatics,
University Medical Center Mannheim
- ² Federated Information Systems,
German Cancer Research Center

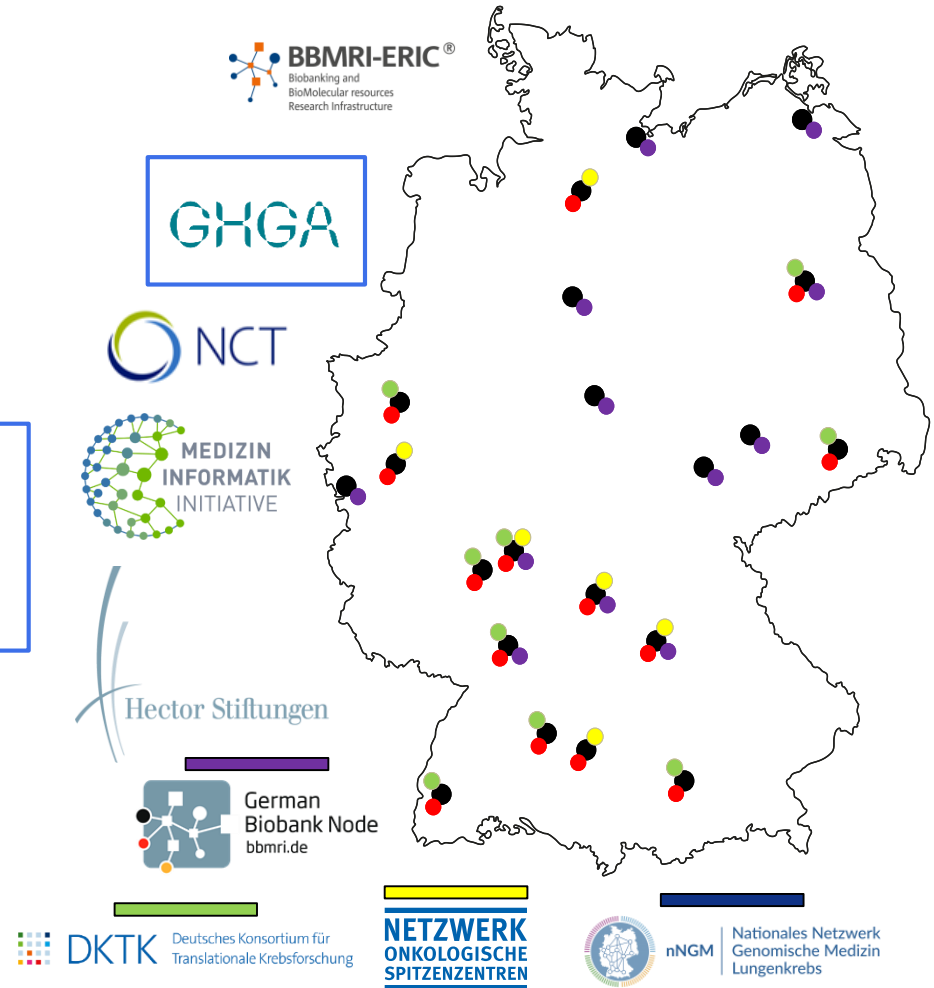


Federated Information Systems (E260)

Leitung: Martin Lablans

Mission: das Bauen von Brücken zwischen DKFZ und Partnern, um Datennetzwerke in Forschung und Versorgung zu ermöglichen.

Privacy-Preserving Linkage von genomischen und klinischen Datensätzen



Anonymisierung vs. Pseudonymisierung

Anonymisiert

Age	Sex	Analysis Cohort	Biopsy Site	Mutation Count	Gen. Alt. Fraction
65	F	BRCA	Brain	169	0.3055
33	F	COLO	Abd. Wall	58	0.4222
27	F	MISC	Cervix	53	0.2452
48	M	COLO	Liver	1174	0.8783
49	F	SKCM	Skin	961	0.6046
44	F	SARC	Liver	16	0.0000
62	F	CHOL	Liver	54	0.1532
47	M	PRAD	Dura	7	0.0041
69	m	SARC	Shoulder	42	0.3390

Beispieldaten aus: Robinson et al. (2017) Integrative clinical genomics of metastatic cancer. *Nature*, 548(7667), 297–303.
https://www.cbioportal.org/study/summary?id=metastatic_solid_tumors_mich_2017

Anonymisierung vs. Pseudonymisierung

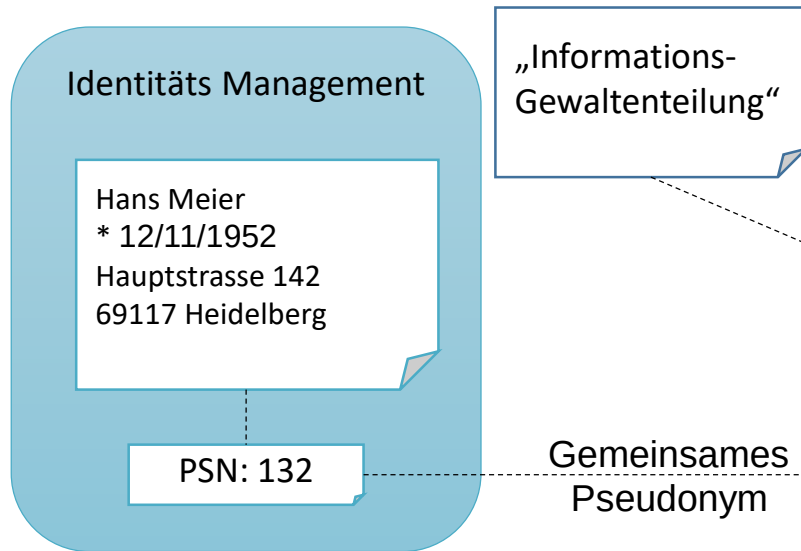
Pseudonymisiert

Patient ID	Age	Sex	Analysis Cohort	Biopsy Site	Mutation Count	Gen. Alt. Fraction
ES_5001	65	F	BRCA	Brain	169	0.3055
ES_5004	33	F	COLO	Abd. Wall	58	0.4222
ES_5005	27	F	MISC	Cervix	53	0.2452
MO_1001	48	M	COLO	Liver	1174	0.8783
MO_1002	49	F	SKCM	Skin	961	0.6046
MO_1005	44	F	SARC	Liver	16	0.0000
MO_1007	62	F	CHOL	Liver	54	0.1532
MO_1008	47	M	PRAD	Dura	7	0.0041
MO_1011	69	m	SARC	Shoulder	42	0.3390

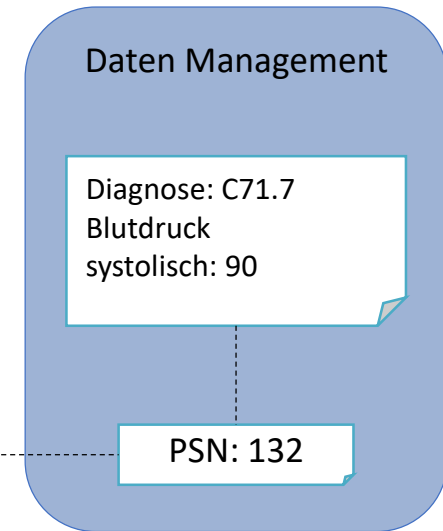
Beispieldaten aus: Robinson et al. (2017) Integrative clinical genomics of metastatic cancer. *Nature*, 548(7667), 297–303.
https://www.cbioportal.org/study/summary?id=metastatic_solid_tumors_mich_2017

Pseudonymisierung in der Medizininformatik

IDAT (Identity Data)



MDAT (Medical Data)



De-Pseudonymisierung



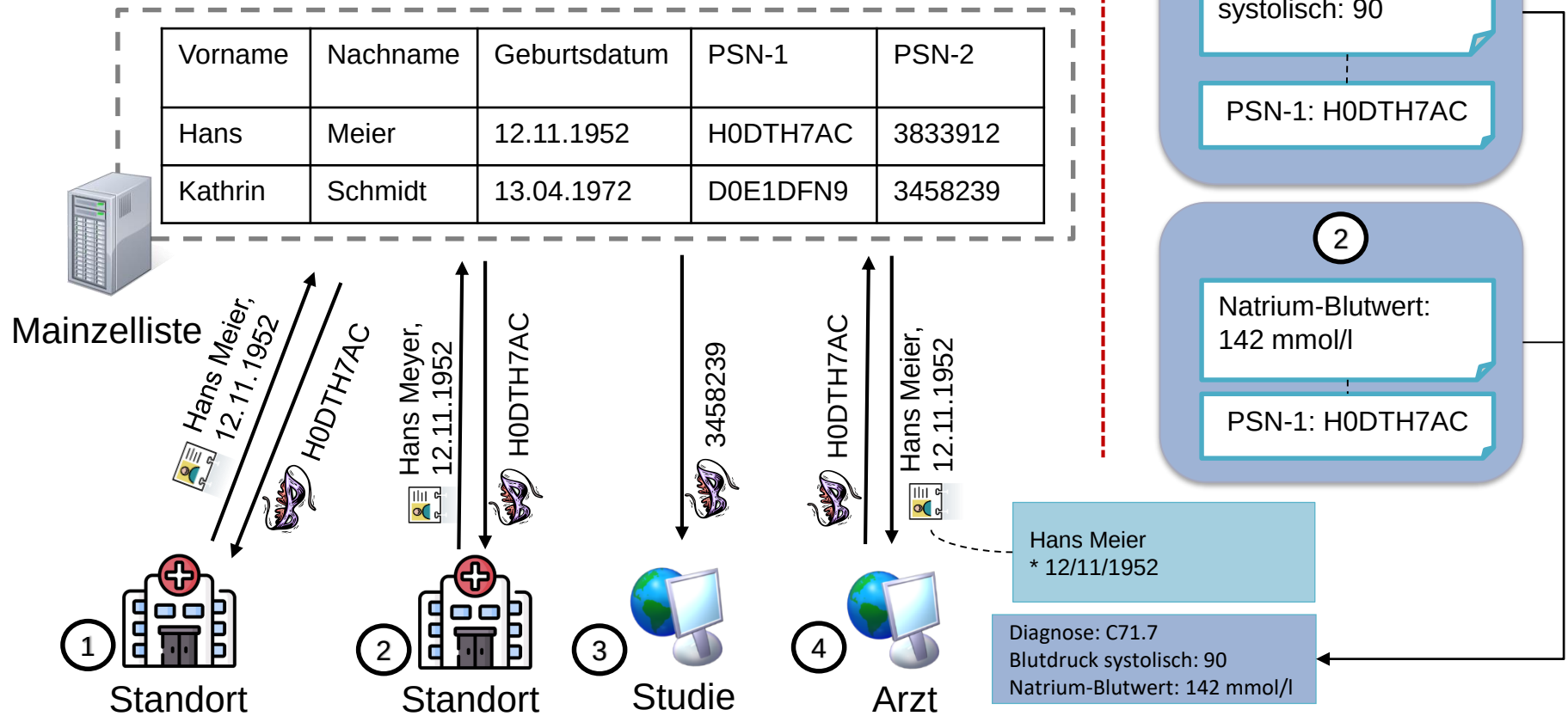
Hans Meier
* 12/11/1952
Hauptstrasse 142
69117 Heidelberg

Diagnose: C71.7
Blutdruck systolisch: 90

ID Management

IDAT (Identity Data)

MDAT (Medical Data)



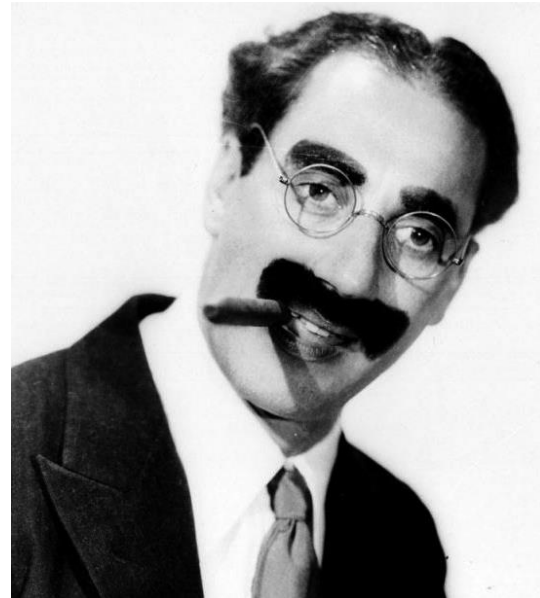
Beispiel Record Linkage (1)

Datenbank A



?
=

Datenbank B



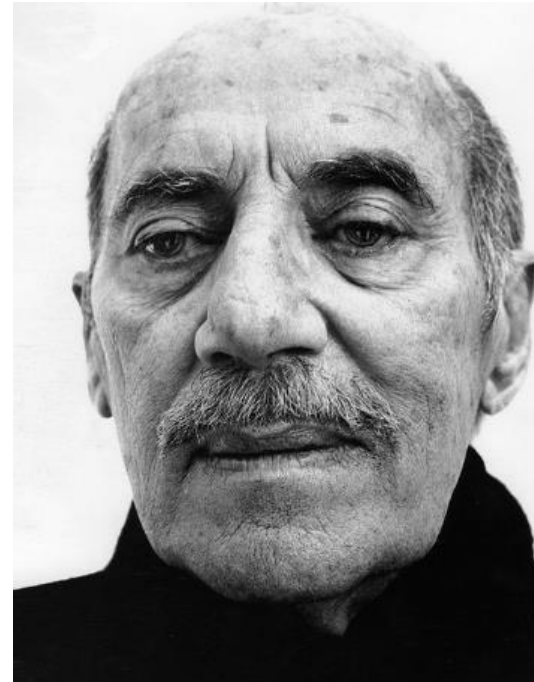
Beispiel Record Linkage (2)

Datenbank A



?
=

Datenbank B



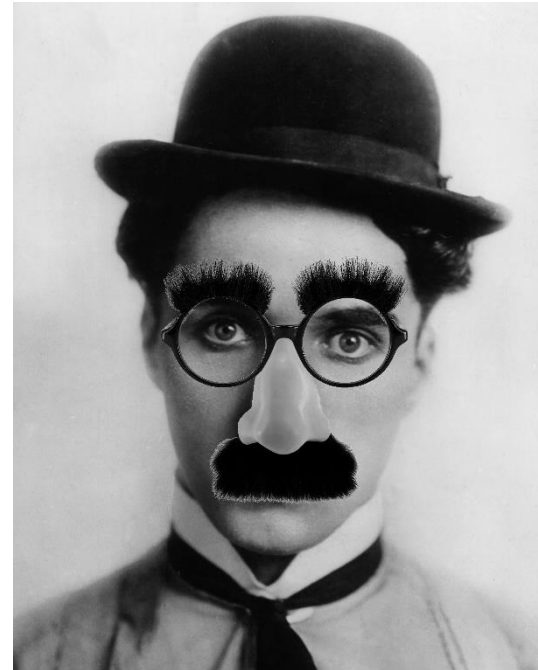
Beispiel Record Linkage (3)

Datenbank A

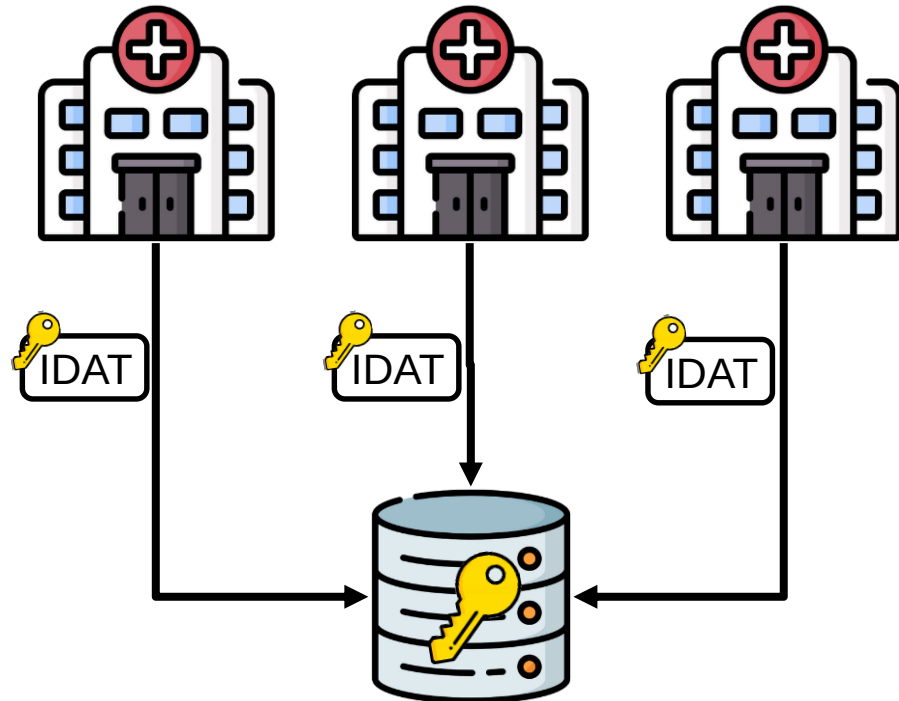
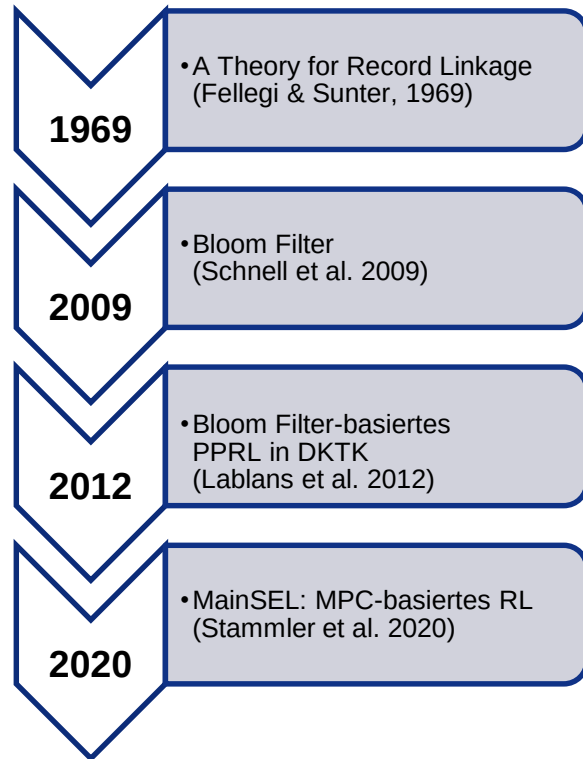


?
=

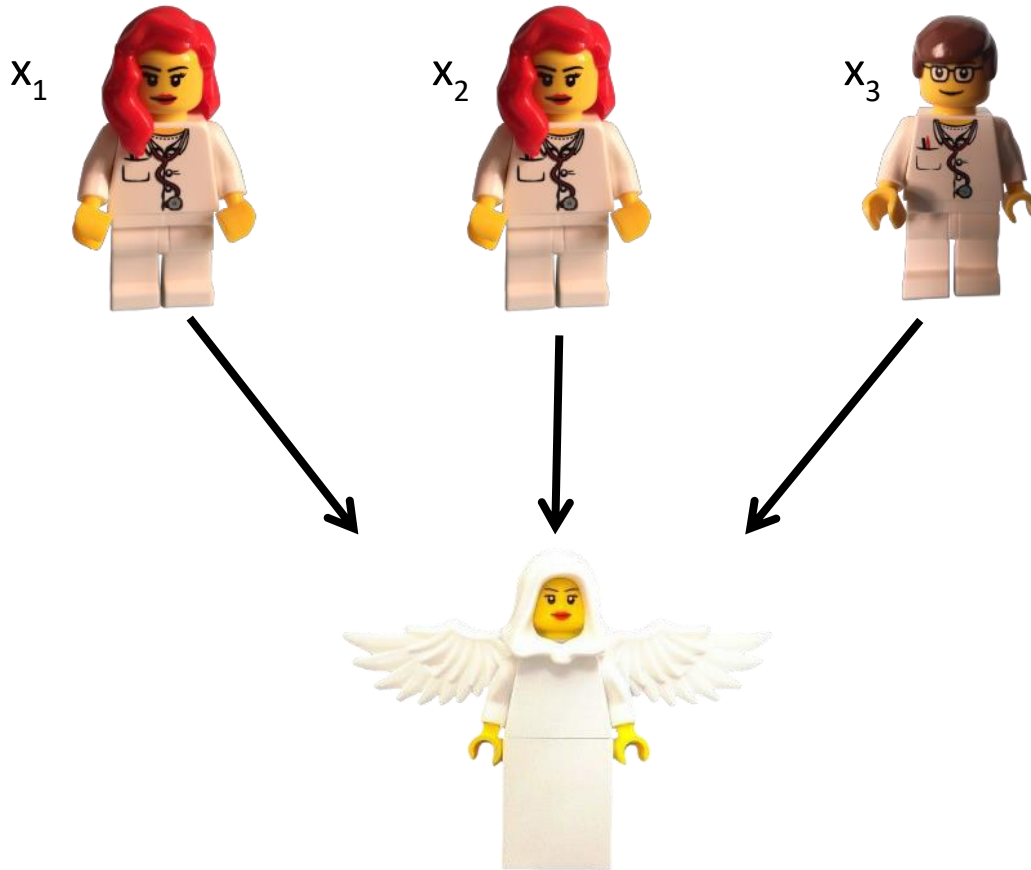
Datenbank B



Eine kurze Geschichte des PPRL

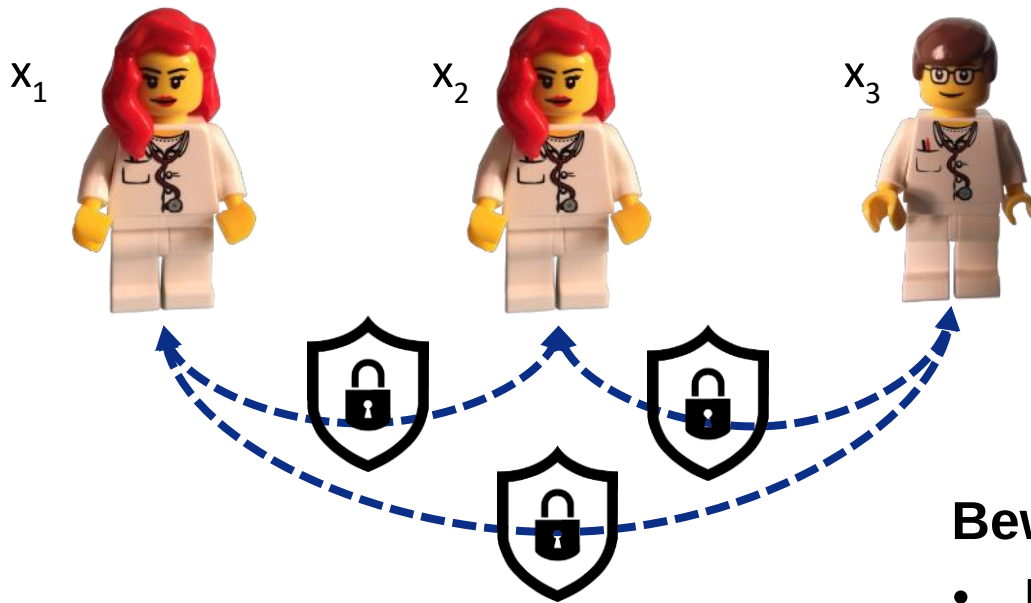


„Klassische“ verteilte Berechnung



Trusted Party
 $f(x_1, x_2, x_3)$

Secure Multi-Party Computation (MPC)

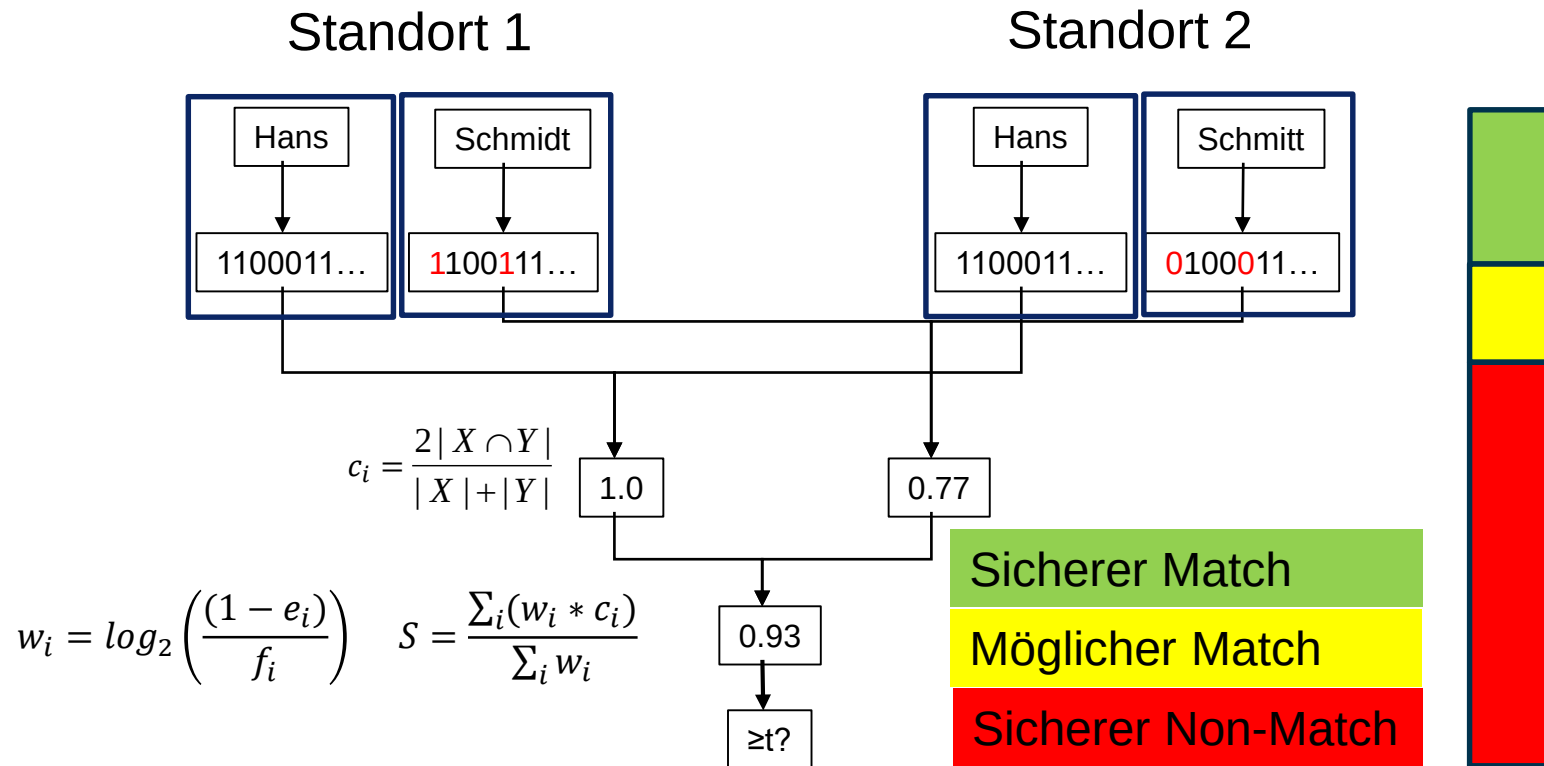


Kryptografisches Protokoll
 $f(x_1, x_2, x_3)$

Beweisbare Garantien u.a.:

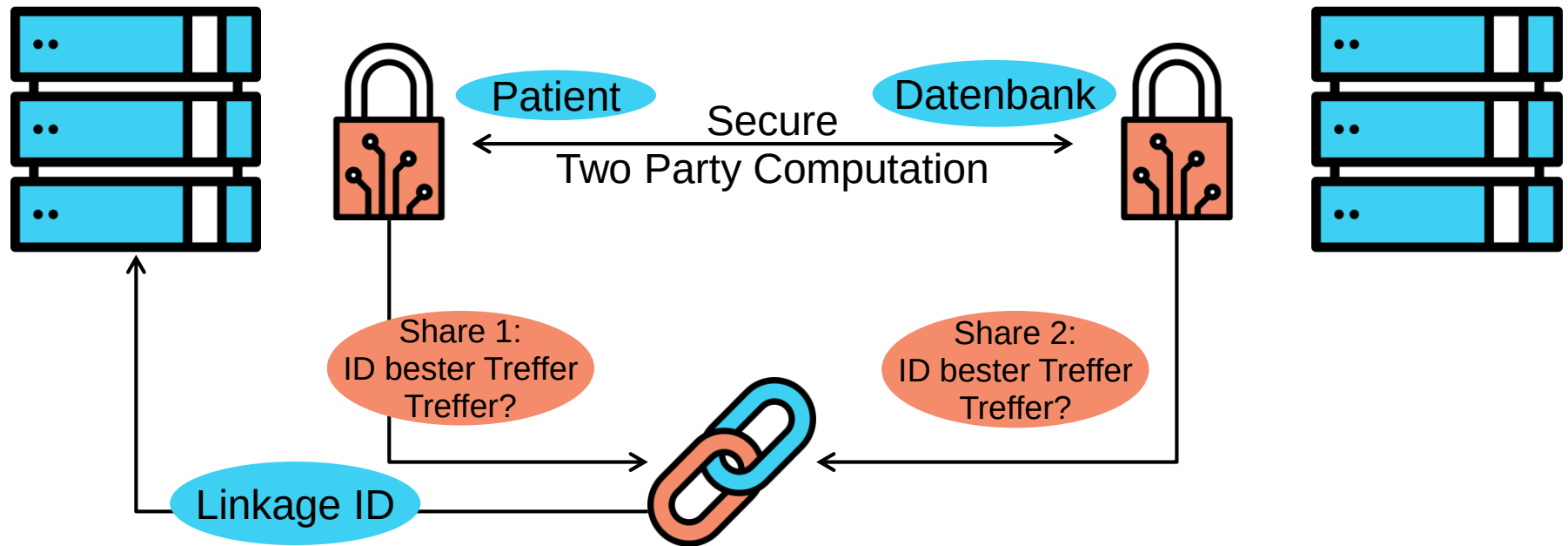
- Privatheit der Eingabedaten
- Korrektheit der Berechnung
- Fairness der Ausgabe

Record Linkage – EpiLink Algorithmus [CTM+05]



P. Contiero et al., The EpiLink record linkage software, in: Methods of Information in Medicine 2005, 44 (1), 66–71.

MPC Record Linkage mit MainSEL [SKS+20]

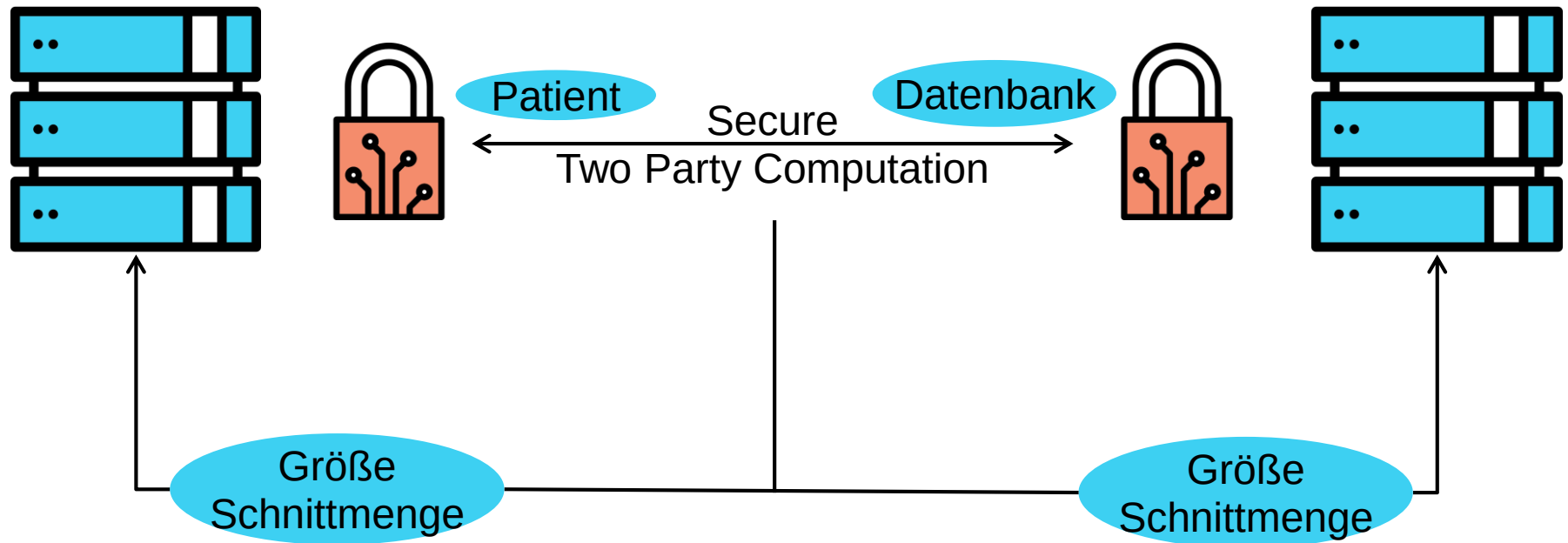


Linkage Service setzt Shares zu Ergebnis zusammen.

- Treffer: Linkage ID = verschlüsselte ID bester Treffer
- Kein Treffer: Linkage ID random generiert

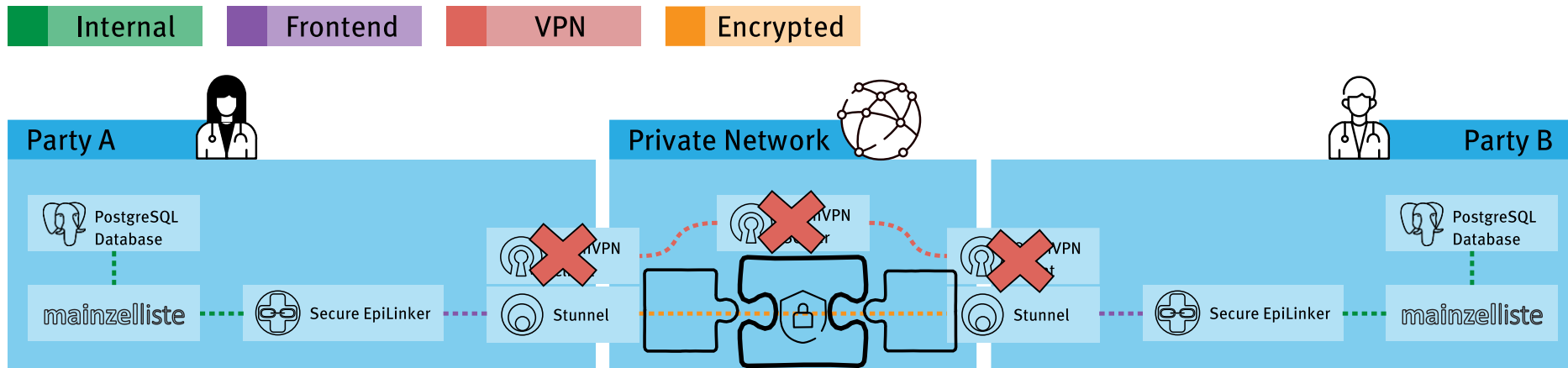
Linkage Service nimmt nicht an der Berechnung teil, sondern erlaubt es das Ergebnis bis zu einer UAC Einwilligung zu „verstecken“

MPC Schnittmenge mit MainSEL [KBT+22]



cord_mi

Deployment MainSEL [KBT+22]



Zusammenfassung

- Wo immer identifizierende Daten genutzt werden, könnte Pseudonymisierung wichtig sein
 - Erzeugen von Studienpseudonymen
 - Rücklieferung von Studienergebnissen
 - Verbindung der Datensätze mehrerer Systeme oder Institutionen
- Record Linkage bezeichnet das Finden von Duplikaten
 - Zur Datenqualitätssicherung
 - Zum Verbinden mehrerer Datensätze
- Mehrere Privatsphäre-schützende Algorithmen existieren, z.B.
 - Bloom-Filter-basiert für große Datenbestände
 - MPC-basiert für höchste Datensicherheit und Dezentralität

Vergleich Bloom Filter und MPC

Bloom Filter

- Salz zur Abhärtung gegen bekannte Attacken nötig
- Keine analytische Sicherheitsabschätzung
- (Kodierte) IDAT verlassen Standort (und werden temporär zentral gespeichert)
- Schnell, Millionen von Patienten
- Erprobte Betriebskonzepte, Datenschutzfolgeabschätzungen ...
- Informed Consent benötigt

MPC

- Keine bekannten direkten Angriffe
- Mathematische Beweisbarkeit der Sicherheit
- IDAT verlassen den Standort nicht
- Langsam, Tausende von Patienten
- Für viele Datenschützer ungewohnt
- Möglicherweise kein Consent für föderierte Analyse benötigt?

Vielen Dank für Ihre Aufmerksamkeit

Wo können Sie Pseudonymisierung nutzen?

Kontakt:

Prof. Dr. Martin Lablans

m.lablans@dkfz.de

Dr. Tobias Kussel

tobias.kussel@dkfz.de

Galina Tremper

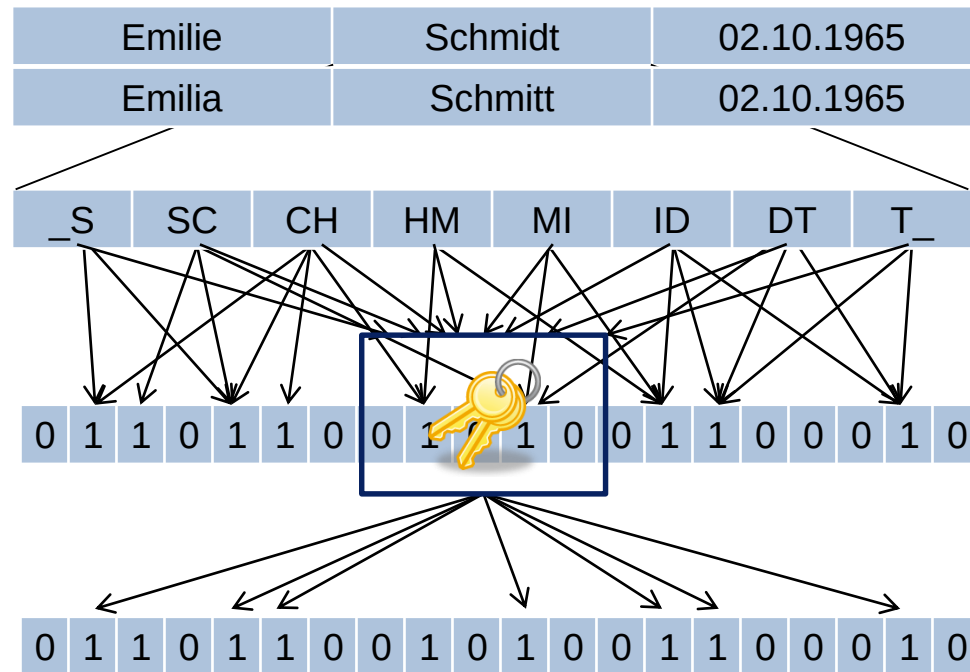
g.tremper@dkfz.de

Referenzen und Quellen

Quellenangaben

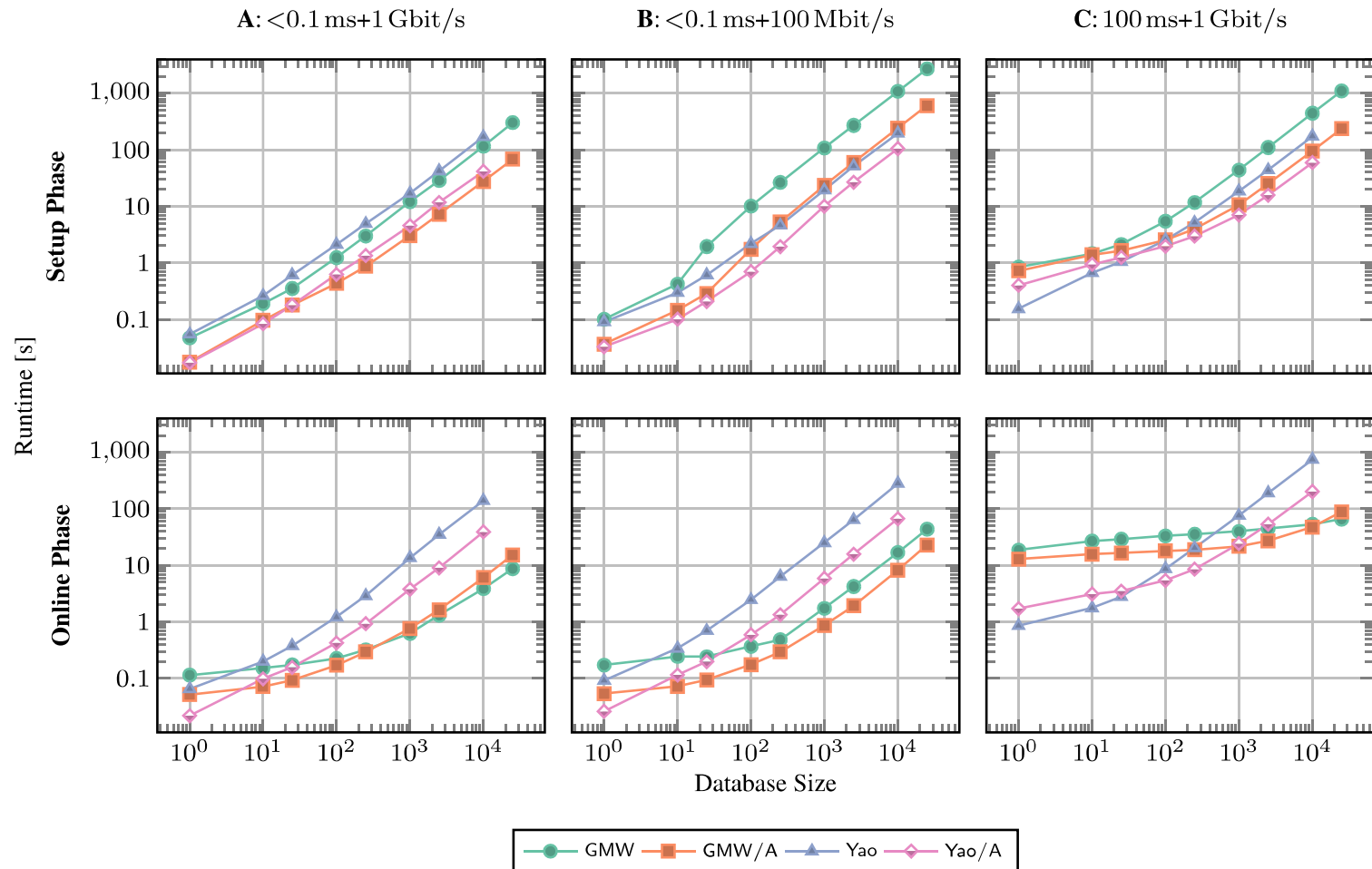
- Alice und Bob Lego-Figuren von Sergio Santos (CC-BY 2.0 Lizenz)
- [CTM+05]: Contiero et al. „The EpiLink Record Linkage Software“ In: *Methods of Information in Medicine*. 2005
- [FS69]: Fellegi und Sunter „A Theory for Record Linkage“ In: *Journal of the American Statistical Association*. 1969
- [KBT+22]: Kussel et al. „Record Linkage based Patient Intersection Cardinality for Rare Disease Studies using Mainzelliste and Secure Multi-Party Computation“ In: *Journal of Translational Medicine*. 2022
- [LSÜ18]: Lablans, Schmidt, Ückert „An Architecture for Translational Cancer Research As Exemplified by the German Cancer Consortium“ In: *JCO Clinical Cancer Informatics*. 2018
- [SBR09]: Schnell, Bachteler und Reiher „Privacy-Preserving Record Linkage using Bloom Filters“ In: *BMC Medical Informatics and Decision Making*. 2009
- [SKS+20]: Stammler et al. “Mainzelliste SecureEpiLinker (MainSEL): Privacy-Preserving Record Linkage using Secure Multi-Party Computation” In: *Bioinformatics*. 2020

(Privacy-Preserving) Record Linkage mit Bloom Filtern



Rainer Schnell, Tobias Bachteler and Jörg Reiher: Privacy-preserving record linkage using Bloom filters. *BMC Medical Informatics and Decision Making* 2009, **9**:41

Performance MainSEL [SKS+20]



Mainzelliste – Product for Community

References

Mainzelliste is used in various medical joint research projects, including:

- German Mukoviszidose Register (Data protection concept)
- European chILD-EU register (Ethics/Data Safety)
- German Cancer Consortium
- Cluster for Individualized Immune Intervention (CI3) (Meeting abstract on IT concept)
- Studies conducted by the LASER group
- The MIRACUM consortium
- The NEW ESID online database network
- The VHL-Register
- CURENet Network researching congenital urorectal malformations
- athesis longlife - das BQS Register
- The ParaReg Registry for lifelong monitoring of paraplegic patients
- The DISCOVER study

Another important use case is pseudonymization in central biobanks, for example:

- Comprehensive Biomaterial Bank Marburg
- Hannover Unified Biobank
- The biobank network of the Ruhr University Bochum (BioNetRUB)

The Mainzelliste API has been implemented in the following projects and software products:

- OSSE – Open Source Registry System for Rare Diseases in the EU
- OpenClinica (see presentation on integrating Mainzelliste and other software)
- secuTrial (see modules description)
- Semantic Clinical Registry System for Rare Diseases
- MOSAIC (the external interface of the "Trusted Third Party Dispatcher" is oriented towards the t
- German Center for Cardiovascular Disease (DZHK) (see MOSAIC and the data protection concep
- German National Cohort (see MOSAIC and the data protection concept)
- Electronic data capture system by Fraunhofer FOKUS
- CentraXX by Kairos GmbH
- Platform for medical research by Climedio Health GmbH
- The NEW ESID online database network
- MaganaMed's trial management software, using its "Mainzelliste Module" for separating patient
- The MainzelHandler open-source library by Daniel Preciado-Marquez et al. for integrating the M
- The Karnak open-source DICOM gateway for data de-identification and DICOM attribute norma
- Elpax by Axis Software and the Gesundes Kinzigtal GmbH.

We have compiled this list from the results of public search engines. If you use the Mainzelliste or its A

Ein Produkt der Community

Riegel J¹, Ben Amor M², Brenner T², Drepper J³, Franke M⁴, Grün M⁵, Hamacher K⁶, Hund H⁷, Knopp C⁸, Kussel T⁶, Lemmer M⁵, Parciak M⁹, Rahm E⁴, Rohde F⁴, Sax U⁹, Schepers J¹⁰, Sehlitz Z⁴, Suhr M⁹, Panholzer T¹, Lablans M²

Föderiertes Record Linkage (Bloomfilter)

Föderiertes Record Linkage auf Basis von Bloomfiltern beherrscht die Mainzelliste bereits seit mehreren Jahren. Seitdem betreiben die Partner Frankfurt und Mainz ein föderiertes Identitätsmanagement im Deutschen Konsortium für Translationale Krebsforschung (DKTK) für Pseudonymisierung von hunderttausenden onkologischen Patienten.

UNIVERSITÄTSGÖTTINGEN, DKTK, UNIVERSITÄT KLINIKUM MANNHEIM

Föderiertes Record Linkage (SMPC)

Der Mainzelliste Secure EpiLinker (MainSEL) erlaubt Record Linkage zwischen mehreren Mainzellisten, ohne dabei die IDAT der Gegenseite zu offenbaren. Diese noch experimentelle Technik bringt ein stark erhöhtes Sicherheitsniveau gegenüber klassischen Techniken des Privacy-Preserving Record Linkage (PPRL) und soll erstmals im MII-Anwendungsfall Cooperation on Rare Diseases (CORD) zum Einsatz kommen.

TECHNISCHE UNIVERSITÄT DARMSTADT, HIGHmed, DKTK, UNIVERSITÄTSGÖTTINGEN

Blocking und Locality-Sensitive Hashing

Im Rahmen der „FASTML“-Förderung des TMF e.V. wurde das Bloomfilter-Matching der Mainzelliste dramatisch beschleunigt. Zur Anwendung kommt hierfür Locality-Sensitive Hashing (LSH [1]). Eine Publikation (zurzeit im Review) belegt Geschwindigkeit und Mächtigke.

UNIVERSITÄT LEIPZIG, TMF

Audit Trail für klinische Studien

Zur Erfüllung von Anforderungen im Bereich klinischer Studien wurde die Mainzelliste um eine Audit Trail-Funktionalität erweitert.

Philippe Universität Marburg, miracum

Handling von eGK-Nummern

An der Universitätsmedizin Göttingen wird der unveränderliche Teil der GKV-Versichertennummer in das Record Linkage einbezogen. Hierfür wurde ein neuer ID-Typ mit entsprechender Validierung in die Mainzelliste integriert.

UNIVERSITÄTSGÖTTINGEN, UMG, HIGHmed

MainzellLibrary

Um die Funktionalität der Mainzelliste für föderiertes Record Linkage in anderen Projekten (u.a. in HIGHmed) zu verwenden, ist eine Umwidmung eines Teils der Mainzelliste erforderlich. Dafür wird die Bibliothek „MainzellLibrary“ geschaffen und unter Apache 2.0 Lizenz veröffentlicht. Zusammen mit den Entwicklern der HIGHmed Data Sharing Frameworks wird eine Weiterentwicklung dieser Bibliothek angestrebt.

HOCHSCHULE WÜRZBURG, HIGHmed

Continuous Integration

Die Integration einer Vielzahl von Codebeiträgen Dritter in die Mainzelliste stellt besondere Anforderungen an Softwarequalität. Hierzu wurden automatisierte Integrationstests geschrieben und in die öffentlich zugreifbare Mainzelliste-CI-Pipeline integriert.

DKTK, UNIVERSITÄTSGÖTTINGEN

Docker-Image

In einem Testlauf der Mainzelliste durch MIRACUM-Standorte 2018/2019 wurde vielfach eine Verfügbarkeit der Software als Docker-Image gefordert und wurde durch Mainz umgesetzt. Das Docker-Image wird mittlerweile automatisch aus der CI-Pipeline (siehe links) gebaut.

UNIVERSITÄTSGÖTTINGEN, miracum

GCP-Validierung

Zum Einsatz in klinischen Studien wird ein Validierungskonzept der Mainzelliste nach GCP erstellt. Die TMF-Arbeitsgruppe IT-Infrastruktur und Qualitätsmanagement (ITQM) hat dieses Vorhaben bereits besprochen und zur Erstellung eines TMF-Projektantrags aufgefordert.

Philippe Universität Marburg, TMF

Gemeinschaftliche Entwicklung in Zahlen

Open Source seit 2013	18 Mitentwickler aus 7 Einrichtungen
41 externe Pullrequests	450 Commits seit 01/2019

1 Medizinische Informatik, Institut für Medizinische Biometrie, Epidemiologie und Informatik, Universitätsmedizin Mainz | 2 Verbundinformationssysteme, Deutsches Krebsforschungszentrum Heidelberg | 3 MII-Koordinierungsstelle, TMF e.V., Berlin | 4 Institut für Informatik, Universität Leipzig | 5 Koordinierungszentrum für Klinische Studien (KKS), Universität Marburg | 6 Computational Biology & Simulation Group, Technische Universität Darmstadt | 7 GECKO Institut, Hochschule Heilbronn | 8 Institut für Translationale Bioinformatik, Universitätsklinikum Tübingen | 9 Institut für Medizinische Informatik, Universitätsmedizin Göttingen | 10 Berlin Institute of Health und Charité – Universitätsmedizin Berlin

Where might you need pseudonymization?

Mainzelliste Dev Team:

- m.benamor@dkfz-heidelberg.de
- g.tremper@dkfz-heidelberg.de
- t.brenner@dkfz-heidelberg.de

More Information:



<https://www.toolpool-gesundheitsforschung.de/produkte/mainzelliste>



<https://hub.docker.com/r/medicalinformatics/mainzelliste/>



<https://www.mainzelliste.de>