



DKTK Deutsches Konsortium für
Translationale Krebsforschung

Use Case 7:

Standortübergreifendes Record Linkage im Deutschen Konsortium für Translationale Krebsforschung (DKTK)

G. Tremper, M. Lambarki, J. Kern, M. Lablans

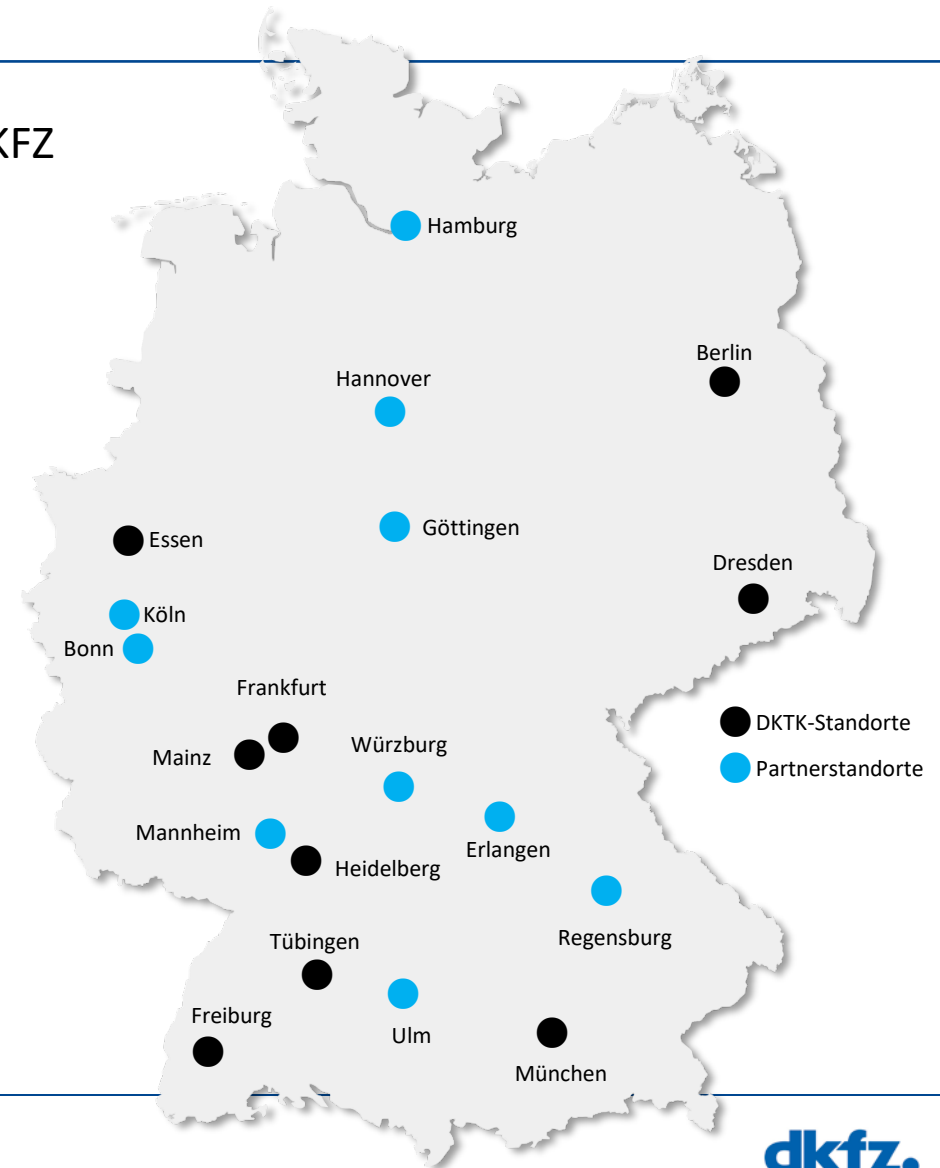
Berlin, 24. Mai 2022

¹Komplexe Datenverarbeitung in der Medizinischen Informatik, Universitätsmedizin Mannheim

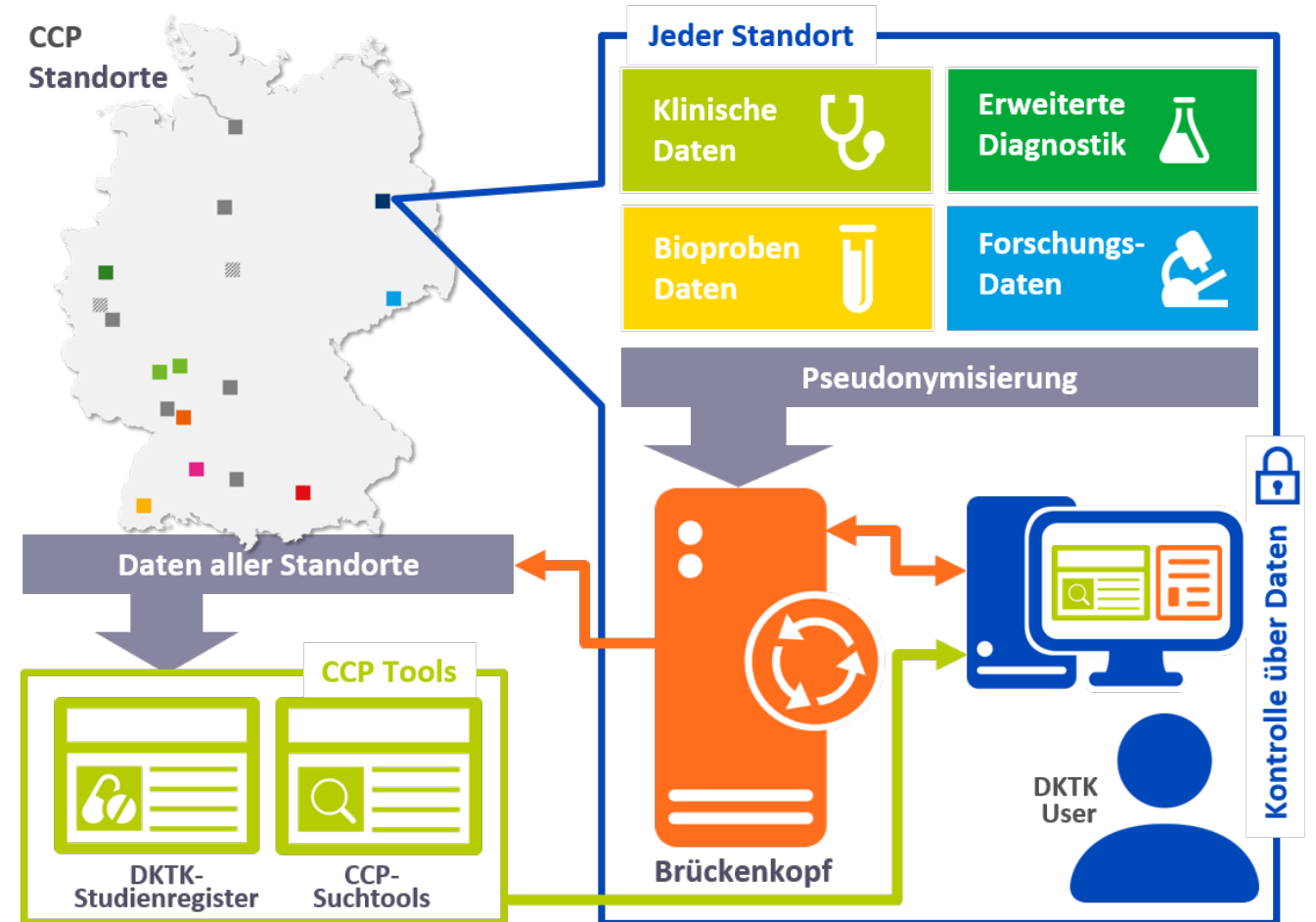
²Verbundinformationssysteme, Deutsches Krebsforschungszentrum

Deutsches Konsortium für Translationale Krebsforschung (DKTK)

- Initiative des BMBF, der teilnehmenden Bundesländer mit DKFZ als Kernzentrum
- Entdeckung, Entwicklung, Erprobung und Anwendung neuer Strategien für die personalisierte Onkologie
- Förderung interdisziplinärer Forschungsthemen an der Schnittstelle zwischen Grundlagenforschung und Klinik, klinischen Studien von innovativen Therapie- und Diagnoseverfahren

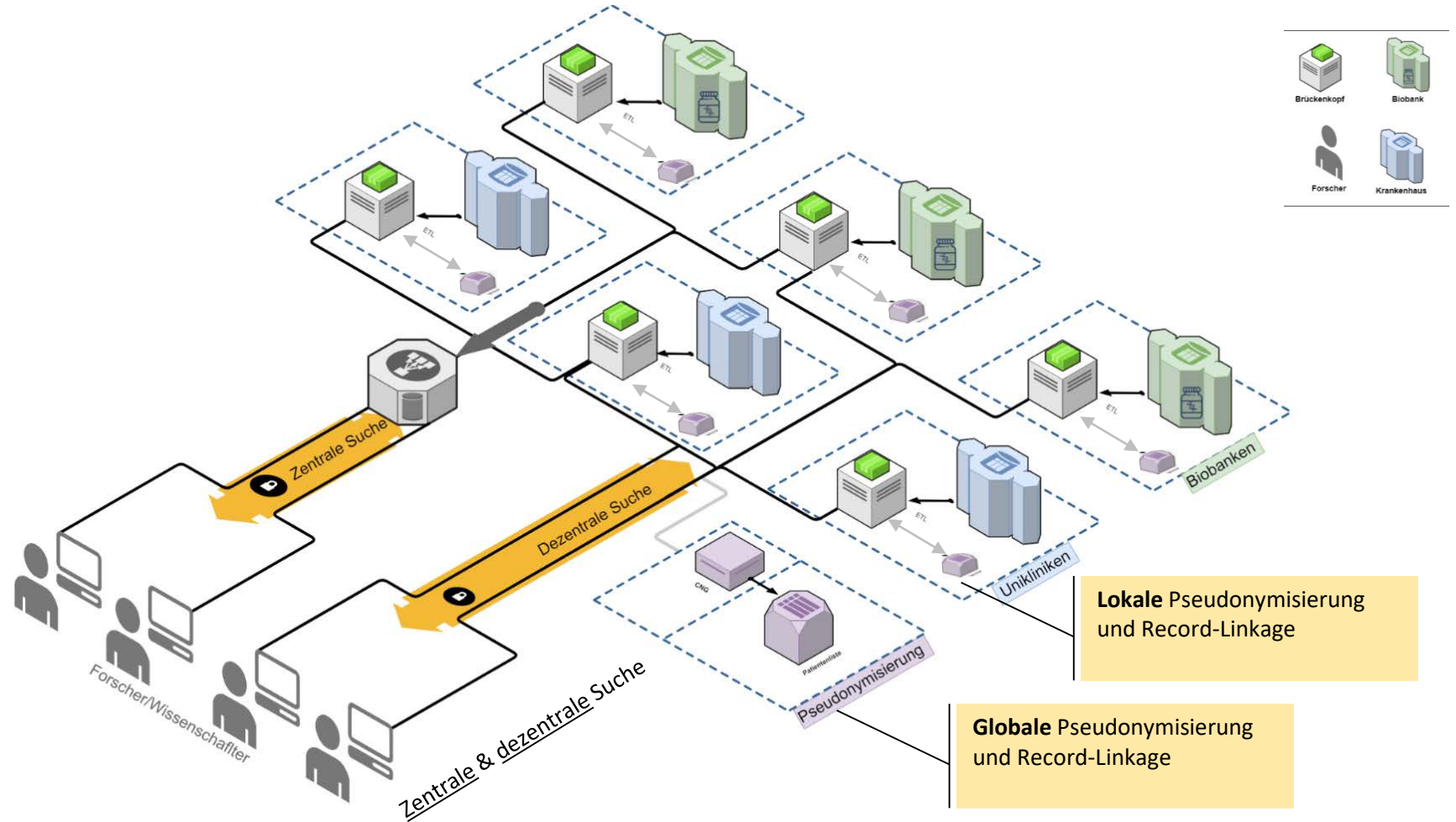


- Informationsdrehscheibe für translationale Krebsforschung
- Standortübergreifende Verknüpfung, Auffinden & Teilen von Daten und Bioproben aus Routineversorgung & Forschungskontext
- Bereitstellung von harmonisierter IT-Infrastruktur
 - CCP-Tools (zentral oder föderiert)
 - CCP-Brückenkopf (am Standort)

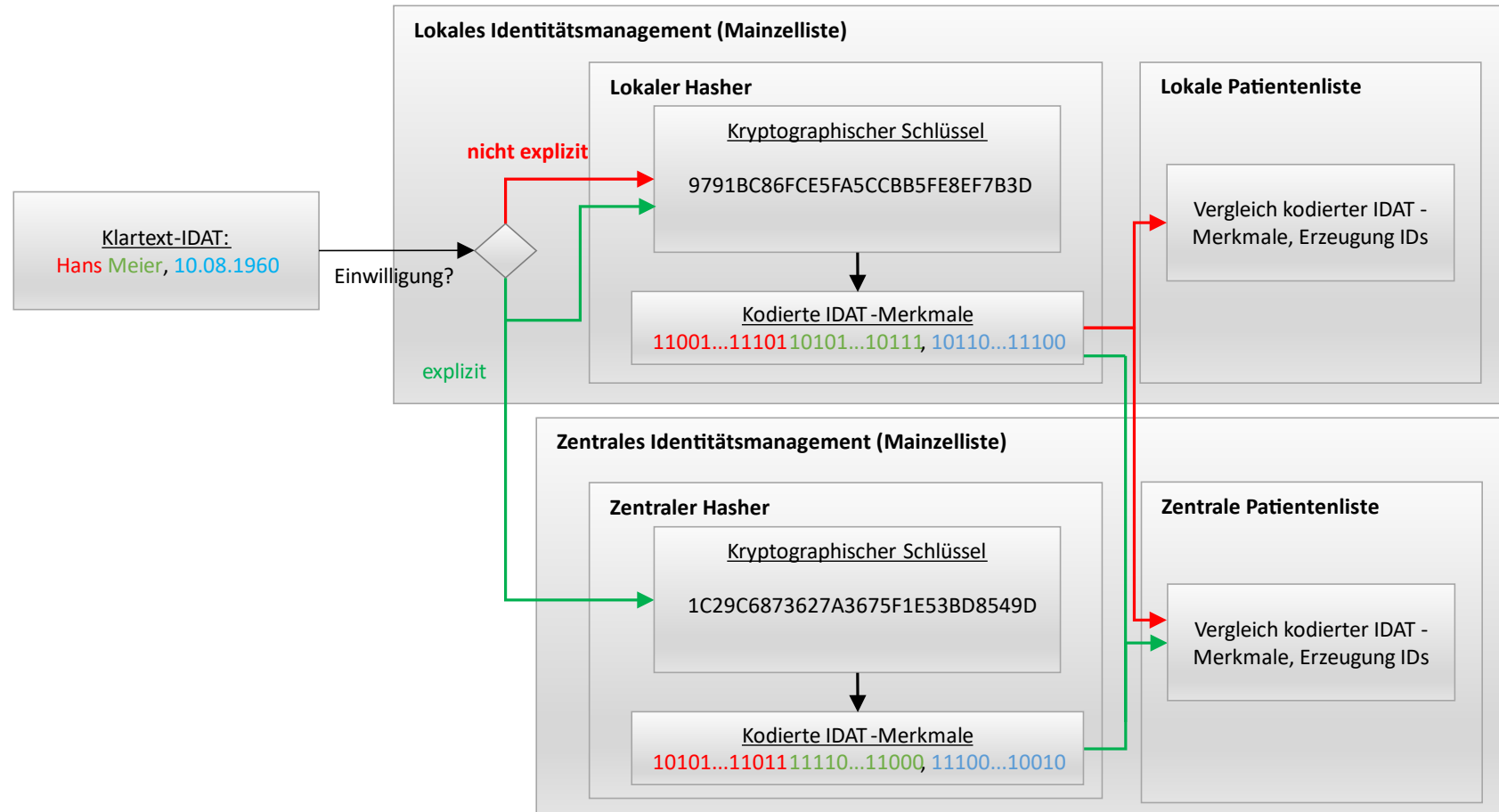


CCP

Klinische Kommunikationsplattform



Privacy-Preserving Record Linkage in DKTK



Record Linkage-Methode

- **Fehlertolerantes** Record Linkage, basiert auf einem **probabilistischen** Algorithmus (EpiLink)
- Verwendete IDAT-Merkmale:
 - ✓ Vorname, Nachname, Frühere Namen, Geburtsdatum, Staatsangehörigkeit, Geschlecht
- Vergleich innerhalb einer Gruppe (Exchange Groups)
- Verwendung der Mainzliste – Record Linkage und Pseudonymisierungssoftware
- Gelebte Open-Source-Software mit regelmäßigen Beiträgen aus vielen weiteren Einrichtungen



mainzliste
pseudonymization and identity management

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, Sehili Z⁴, Suhr M⁹, Panholzer T¹, Lablans M²

Föderiertes Record Linkage (Bloomfilter)

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

UNIVERSITÄTSMEDIZIN GÖTTINGEN | dkfz. DEUTSCHES KONSORTIUM FÜR TRANSLATIONALE KREBSFORSCHUNG | DTKF | German Cancer Consortium | UNIVERSITÄT FRANKFURT AM MAIN | UNIVERSITÄT MAINZ

Föderiertes Record Linkage (SMPC)

Der Mainzliste Secure EpiLinker (MainSEL) erlaubt Record Linkage zwischen mehreren Mainzlisten, 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 | dkfz. DEUTSCHES KONSORTIUM FÜR TRANSLATIONALE KREBSFORSCHUNG | v1.10

Blocking und Locality-Sensitive Hashing

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

UNIVERSITÄT LEIFZIG | TMF | v1.9

Handling von eGK-Nummern

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

UNIVERSITÄTSMEDIZIN GÖTTINGEN | UMG | HiGHmed | v1.9

Continuous Integration

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

dkfz. DEUTSCHES KONSORTIUM FÜR TRANSLATIONALE KREBSFORSCHUNG | v1.9

GCP-Validierung

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

Philipp-Universität Marburg | TMF | Ausblick

Audit Trail für klinische Studien

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

Philipp-Universität Marburg | miracum | v1.9

MainzLibrary

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

H+H HOCHSCHULE HEILBRONN | HiGHmed | v1.10

Docker-Image

In einem Testlauf der Mainzliste 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.

UNIVERSITÄTSMEDIZIN GÖTTINGEN | miracum | v1.8

Gemeinschaftliche Entwicklung in Zahlen

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

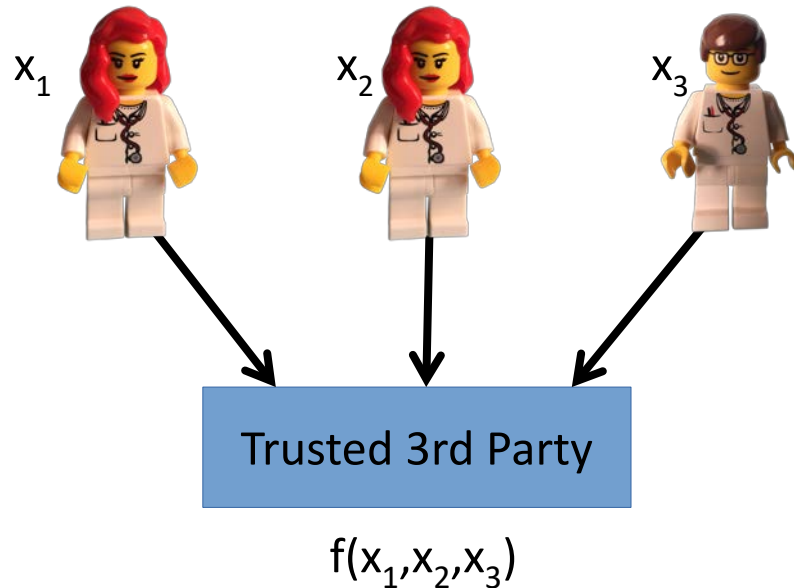
[1] Medizinische Informatik, Institut für Medizinische Biometrie, Epidemiologie und Informatik, Universitätsmedizin Mainz | [2] Verbundinformationssysteme, Deutsches Krebsforschungszentrum Heidelberg | [3] MII-Koordinierungsstelle, TMF e.V., Berlin | [4] Institut für Informatik, Universität Leipzig | [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 and Charité – Universitätsmedizin Berlin

Herausforderungen

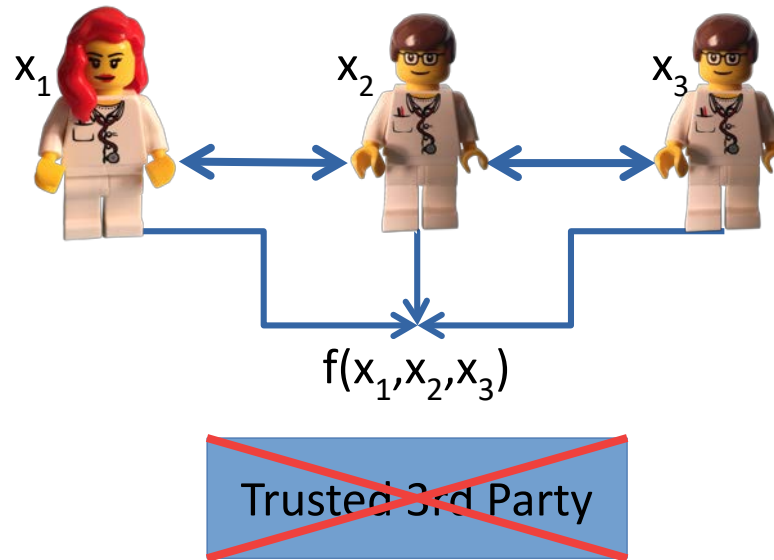
- Technische und personelle Ausstattung der Standorte
 - Minimale technische Anforderungen, Standortbetreuung
 - ➔ Brückenkopfes als einheitliches „Paket“ für alle Standorte minimiert den Aufwand
- Heterogenität der Daten
 - Datenintegration, ETL-Prozesse
 - ➔ Zentral vorbereitete Integrationsstrecken für die wichtigsten Datenquellen (ADT2FHIR für Tumordoku)
- Patienteneinwilligungen
 - In der Praxis rollt das DKTK (begleitet durch das CCP Office) eigene Pat.-Ew. aus
 - ... aber bei weitem nicht für alle Routine-Patienten

Ausblick: Next Generation Privacy Preserving-Record Linkage

Aktueller Stand



SMPC-basiertes Record Linkage [1]



Using MPC as the cryptographic basis of MainSEL provides vastly higher security guarantees than existing solutions, e.g. based on centralized Bloom filters. In particular:

- Sensitive patient data is protected using modern, tried-and-tested cryptographic primitives. Informally speaking, even the “weakest” cryptographic building-block used in MainSEL (OT-Extension) uses state-of-the-art elliptic curve cryptography and is deemed safe until a sufficiently powerful quantum computer is built.
- No trusted third party or central component has to take part in the computation. There is no central collection of sensible identifying data of any kind.
- Even if all but one participating parties are compromised, all uncompromised parties’ patient data remain private.

aus: [2]

Kann man hier überhaupt noch von Datenübermittlung sprechen?

[1] Stammler et al, Mainzelliste SecureEpiLinker (MainSEL): Privacy-Preserving Record Linkage using Secure Multi-Party Computation, Bioinformatics, 2020,; btaa764, <https://doi.org/10.1093/bioinformatics/btaa764>

[2] Kussel et al, Record Linkage based Patient Intersection Cardinality for Rare Disease Studies using Mainzelliste and Secure Multi-Party Computation. In review, full preprint available at <https://doi.org/10.21203/rs.3.rs-1486673/v1>

Fragen?

Weitere Informationen

- **DKTK:** Deutsches Konsortium für Translationale Krebsforschung
 - <https://dktk.dkfz.de/>
- **CCP:** Die Klinische Kommunikationsplattform:
 - <https://dktk.dkfz.de/klinische-plattformen/ueber-die-ccp/about-ccp>
- **Mainzliste:**
 - <https://www.mainzliste.de>



DKTK

Deutsches Konsortium für
Translationale Krebsforschung

Use Case 6:

Linkage von klinischen Routine- und Studiendaten über mehrere rechtliche Träger hinweg am Beispiel des nationalen Netzwerks Genomische Medizin Lungenkrebs

G. Tremper, M. Lambarki, J. Kern, M. Lablans

Berlin, 24. Mai 2022

¹Komplexe Datenverarbeitung in der Medizinischen Informatik, Universitätsmedizin Mannheim

²Verbundinformationssysteme, Deutsches Krebsforschungszentrum

dkfz.

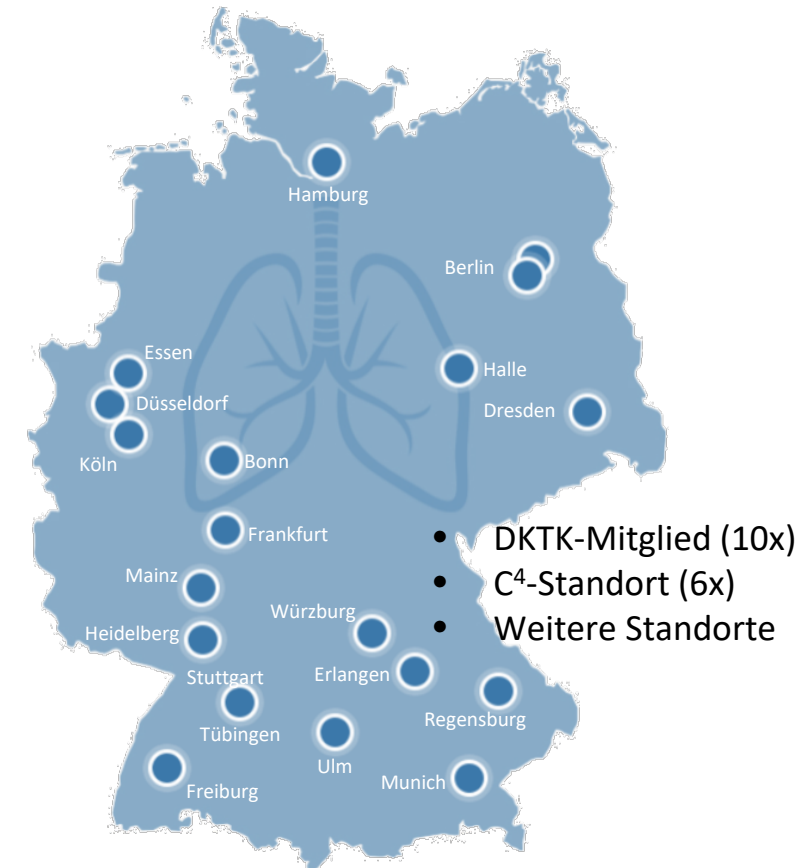
Deutsches Konsortium für
Translationale Krebsforschung

nNGM: nationales Netzwerk Genomische Medizin (Lungenkrebs)

Zusammenschluss von **20** Krebszentren plus ihre regionalen Netzwerke

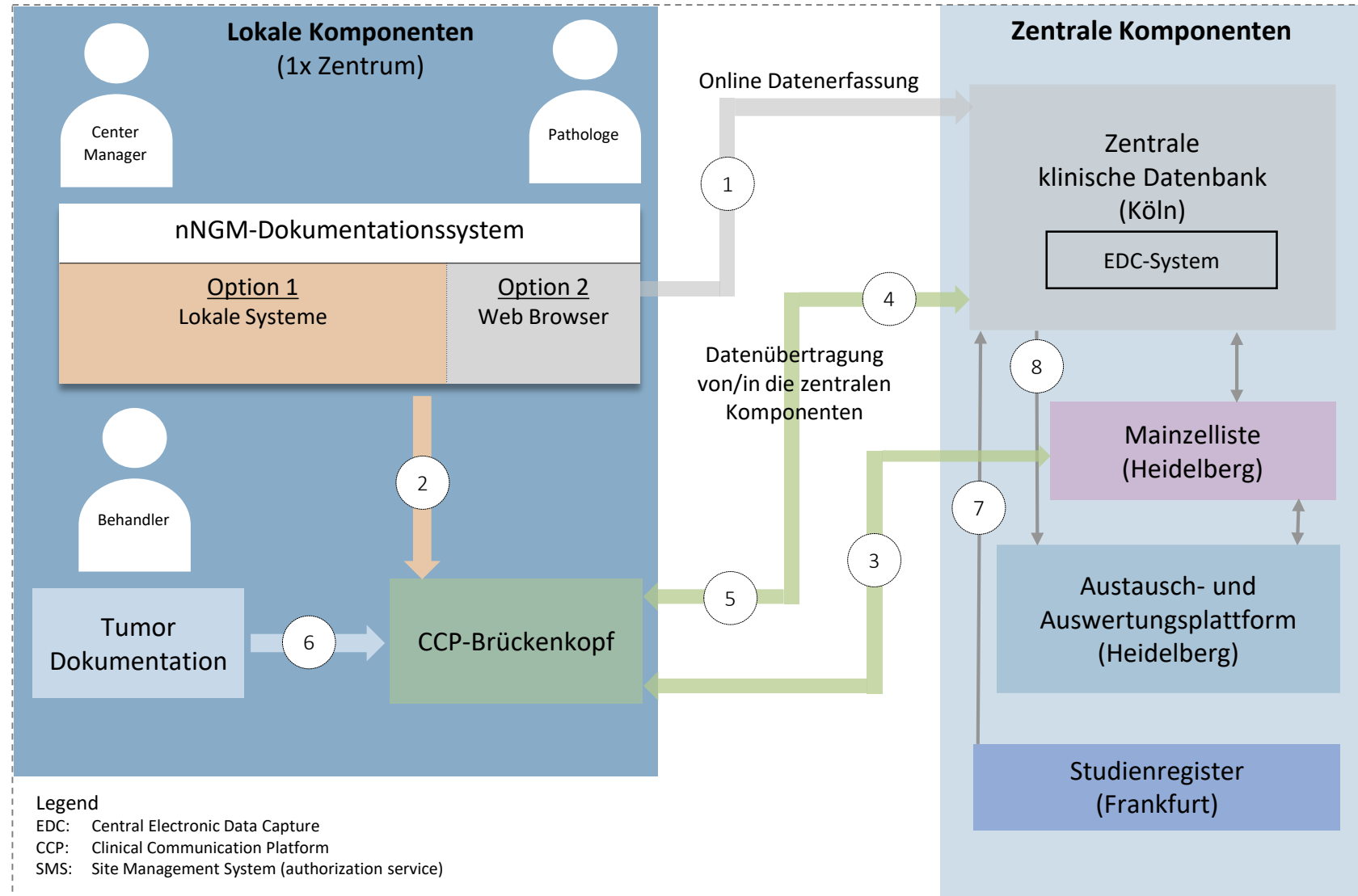
Hauptziele:

- **Für Patienten:** Zugang zu molekularer Diagnostik
→ Identifikation genetischer Veränderungen in den Tumorzellen
- **Für Patienten:** Zugang zu personalisierten Therapien
- Stärkung der **Krebsforschung**, insb. durch eine gemeinsame Datenbasis zwischen den Standorten



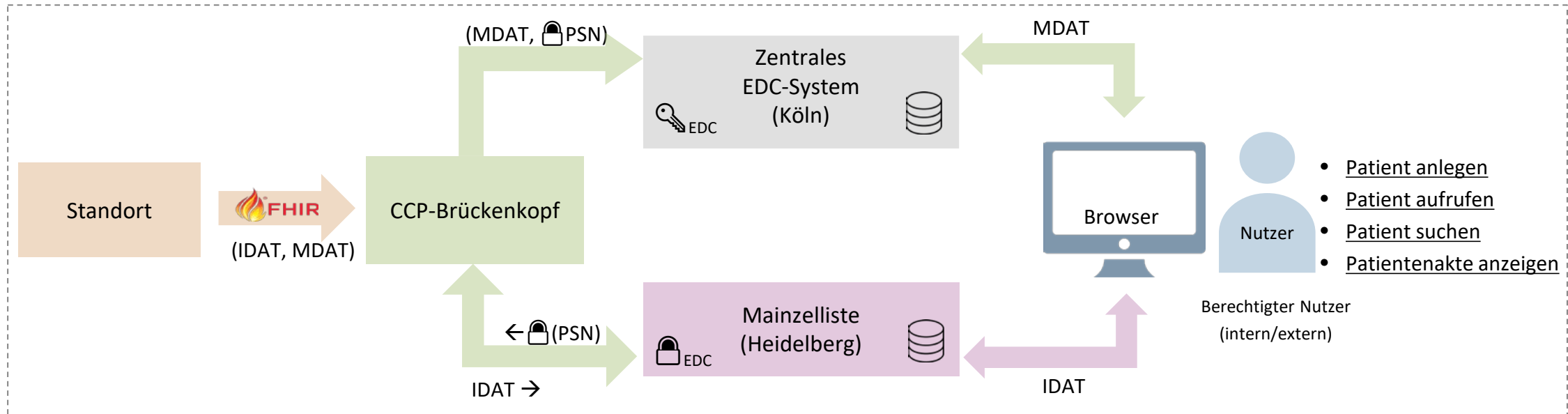
Datenfluss (1/2)

- Föderierte Architektur mit lokalen und zentralen IT-Komponenten
- Standort: direkte oder indirekte Dokumentation
- EDC-System darf nur Pseudonyme speichern
- IDAT werden über die Mainzliste pseudonymisiert und Record-Linkage durchgeführt
- Ergebnisse der Diagnostik sowie Therapieoptionen werden im EDC-System bereitgestellt.
- Mutationsbezogen können klinische Studien aus dem Studienregister identifiziert und empfohlen werden
- **Für die Forschung:**
 - Download in den Brückenkopf
 - Import in eine zentrale Auswertungsplattform



Datenfluss (2/2)

Nachladen der IDAT in den Browser der behandelnden Ärzte



Record Linkage-Methode

- **Analog zu DKTK**
 - *Algorithmus*: Probabilistischer Algorithmus (EpiLink)
 - *Mainzliste*: Record Linkage und Pseudonymisierung
 - Projektbezogene (nNGM) Beschreibung der Pseudonymisierungsschritte
- **Unterschiede/Besonderheiten:**
 - IDAT werden erhoben (→ Nachladen im Behandlungskontext)
 - Record-Linkage nutzt direkt die IDAT (statt kodierter Merkmale wie in PPRL)
 - Versichertennummer wird erhoben und für Plausibilitätsprüfungen genutzt
 - Jedes System (EDC, Brückenkopf...) oder Nutzergruppe (Behandler, Forscher) hat ein eigenes Pseudonym
 - Bei Bedarf werden die Schlüssel im Datenfluss „umgeschlüsselt“ (z.B. Standort-ID <-> EDC-ID)
 - Geht ein Pseudonym über eine Drittstelle wird eine asymmetrische Verschlüsselung angewendet

Herausforderungen

- Record-Linkage/Pseudonymisierung für unterschiedliche Kontexte:
 - Behandlung ↔ Forschung ↔ ...
- Zeitaufwendige Abstimmungsprozesse mit den Datenschützern der Standorte
- Keine einheitliche Krankenversicherungsnummer im privaten Sektor
 - Datenvalidierungsregeln können hierfür nicht angewendet werden
 - Verlust an Datenqualität
- Eingeschränkte personelle Ressourcen am Standort, Verfügbarkeit der IT-Systeme
 - Nutzung der bereits aufgebauten DKTK/CCP-Brückenköpfe vermeidet Extra-Aufwand (→ UC 7)
- Datenzusammenführung/-trennung, Löschung falscher IDAT/Verknüpfungen...
- Klärung der Konflikte bei unterschiedlichen IDAT für denselben Patienten
 - Wer hat Recht?



DKTK

Deutsches Konsortium für
Translationale Krebsforschung

Vielen Dank!