Pilotprojekt genomDE

Konzept für Sichere Verarbeitungs-umgebungen im Modellvorhaben Genomsequenzierung (§64e SGB V)

Prof. Dr. Thomas Berlage (Fraunhofer FIT), Dr. Knut Kaulke (TMF), Prof. Dr. Sebastian Graf von Kielmansegg (Universität Kiel), Prof. Dr. Oliver Kohlbacher (Universität Tübingen), Dr. Pascal Kraft (GHGA), Prof. Dr. Michael Krawczak (TMF, Universität Kiel), Friedrich von Kessel (TMF)

19. Dezember 2025

Gefördert durch:

Bundesministerium für Gesundheit

aufgrund eines Beschlusses des Deutschen Bundestages

Inhalt

Executive Summary

Instrumente und Organisationseinheiten einer SPE
Die Datennutzung im Modellvorhaben (MV) muss bei allen ihren einzelnen Schritten "Auffinden", "Machbarkeit prüfen", "auf Antrag bereitstellen", "in SPE verarbeiten" und "Ergebnisse abschließen" kundenorientiert unterstützt werden. Die dazu erforderlichen technischen Komponenten stellt überwiegend der Plattformträger bereit (z.B. Nutzeridentifizierung, Antragsportal, Datenbereitstellung, Software Repository) oder zugelassen (zum Beispiel Datendienste). Zusätzliche agieren separat im Rahmen der Datennutzung die Datenknoten (Datenverwaltung) und die Vertrauensstelle (Datenverknüpfung).

Konzept und technischer Demonstrator SPE genomDE
Das vorliegende Konzept sieht einen technischen Demonstrator in einem vereinfachten Szenario vor. Dies ermöglicht die frühe Erlangung von Nutzerfeedback und Erkenntnissen aus dem praktischen Betrieb. Vorgeschlagen wird ein 3-phasiges Betriebsmodell bestehend aus 1. Bereitstellung (Instanziierung und Datenzugriff), 2. Nutzung und 3. Beendigung (Ergebnisausleitung, Rückführung dafür vorgesehener Nutzungsergebnisse, Deprovisionierung). Die Klärung des Föderierungsbedarfs sowie der rückzuführender Daten erfolgt im Antragsprozess. Die projektbezogene Pseudonymisierung kann an die SPE-Betreiber delegiert werden, um eine doppelte Datenhaltung zu vermeiden.

Use Cases der Datennutzung
Aus einer Vielzahl möglicher Use Cases wurden 4 Use Cases ausgewählt, die konkreter betrachtet werden:

- UC1: Datenübersicht (Dashboard)
- UC2: Ähnlichkeitssuche ("Patients-like-mine")
- UC3: Qualitätssicherung + Forschung mit Genomdaten (GRZ)
- UC4: Adaptierbare Umgebung (R, Python)

Ein grundlegendes Architekturmodell erlaubt es, alle vier Use Cases in drei verschiedene Komponenten mit einfachen Schnittstellen abzubilden, aus denen sich alle Netzwerkknoten (ob mit Datenhaltung oder nur für Verarbeitung) zusammensetzen: Datenzonen, Abfragezonen und Analysezonen. Die Zonen sichern Datenhaltung und Datenbereitstellung von der nutzergetriebenen Verarbeitung ab.

Umsetzungsplan
Für die Umsetzung wird empfohlen, zunächst zwei Datendienste (ein Übersichts-Dashboard UC1 und eine Basisversion von "patients-like-mine" UC2) in Verbindung mit einer föderierten Datenbereitstellung und -berechnung zu realisieren. Damit könnte ein unmittelbarer Nutzen der Plattform für die lernende Versorgung ermöglicht werden. Parallel, aber entlang einer etwas längerer Zeitschiene, soll die Nutzung der Genomrechenzentren mit verknüpften klinischen Daten (UC3) über ein Antragsportal umgesetzt werden. Damit würde eine SPE-Funktionalität bereitgestellt, die spezifisch für die Daten des MV ist. Als dritter Schritt sollte dann die Datennutzung über das MV hinaus mit anderen Entwicklungen (insbesondere FDZ und EHDS) integriert werden.

Rahmenbedingungen und Governance
Umfang und Zweck der Nutzung von MV Daten (soweit vom Plattformträger bereitgestellt) sind in §64e SGV und GenDV geregelt. Die rechtliche und organisatorische Verantwortung hierfür liegt beim Plattformträger. Alle geeigneten Leistungserbringer haben ein Nutzungsrecht für Versorgungszwecke, unabhängig von ihrer eigenen Teilnahme am MV. Über die Forschungsnutzung von MV Daten (soweit vom Plattformträger bereitgestellt) entscheidet der im §64e SGB V vorgesehene „wissenschaftliche Beirat". Weitere Nutzungsrechte ergeben sich aus den entsprechenden Einwilligungen der Patienten (z.B. MII Broad Consent, bei dem die Nutzungsrechte beim jeweiligen Versorger liegen).

Die Bereitstellung von MV Daten durch den Plattformträger erfolgt über die gesetzlich dafür vorgesehenen Einrichtungen (Treuhandstelle, GRZ, KDK und Datendienste) und - soweit rechtlich und technisch möglich - in SPEs. Anforderungen an Nutzer bzw. Verarbeiter von MV Daten legt der Plattformträger fest.

Entgelte und Vergütung für die Nutzung der SPE-Dateninfrastruktur

Entgelte sollen grundsätzlich kostenbasiert festgesetzt werden, sind aber mit „Marktpreisen" für ähnliche Leistungen abzuwägen. Ein einheitlicher Entgeltkatalog gibt den potenziellen Nutzern Sicherheit und vereinfacht jegliche Form interner Verrechnungssysteme. Er ermöglicht eine schnelle Kalkulation der projektspezifischen finanziellen Aufwände. Vergleichbare Infrastrukturelemente (z.B. alle KDK) verlangen für vergleichbare Leistungen identische Entgelte. Um diese Preise, und damit auch die Forschungsbudgets der Nutzer, in einem vertretbaren Rahmen zu halten, könnte zu einem späteren Zeitpunkt für die erbrachten Leistungen der Plattform ein Effizienzfaktor für einzelne Infrastrukturelemente vorgesehen werden.

Der Plattformträger entlastet die einzelnen Infrastrukturelemente per zentraler Abrechnung von ihren Kosten eines Datennutzungsprojektes und gibt diese Kosten gemäß Entgeltkatalog an die Forschenden weiter.

1. Einleitung

Das Pilotprojekt genomDE startete am 1. Oktober 2021 und endet am 31. Dezember 2025. Die in genomDE vertretenen genommedizinischen Netzwerke, wissenschaftlichen Institutionen und Patientenorganisationen haben in den vergangenen Jahren ein Konzept für die Dateninfrastruktur und wesentliche Voraussetzungen für das Modellvorhaben Genomsequenzierung (§64e SGB V) entwickelt.

Im nun vorgelegten Konzept für Sichere Verarbeitungsumgebungen (Secure Processing Environment, SPE) im Modellvorhaben Genomsequenzierung werden die Voraussetzungen und das Vorgehen zur sicheren Datennutzung der nationalen genommedizinischen Dateninfrastruktur beschrieben. Ein solches Konzept ist notwendig, um die ab Sommer 2024 im Modellvorhaben gesammelten klinischen und genomischen Daten in angemessener Sicherheit für Versorgungsabfragen (z. B. „patients like mine") und Forschungsvorhaben verfügbar zu machen und damit den Erfolg des Modellvorhabens Genomsequenzierung und der nationalen genommedizinischen Dateninfrastruktur sicherzustellen[1].

In der Arbeitsgruppe 3 „Informatik" im Pilotprojekt genomDE wurde in 2024 die Konzeptarbeit für SPE für das Jahr 2025 vorbereitet. In zwei Workshops im Oktober und November 2024 wurden die Anforderungen der genommedizinischen Forschung und Versorgung an die zukünftige Datennutzung und die Erfahrungen von drei europäischen Ländern mit SPE bei der sicheren Nutzung von genommedizinischen Dateninfrastrukturen erarbeitet. Nach der Mitteilung der Aufstockung des genomDE-Budgets vom 17.02.2025 hat die Task Force SPE[2] mit der Arbeit begonnen. Sie hat in insgesamt 22 Sitzungen das Konzept entwickelt und formuliert. In einem dritten Workshop Ende Juni 2025 wurden die Use Cases noch einmal konkretisiert. Eine Expertise von Prof. Dr. Sebastian Graf von Kielmansegg wurde für das Kapitel 5.2 eingeholt.

Das Konzept wurde in einem gemeinsamen Termin der sechs Arbeitsgruppen im Pilotprojekt genomDE am 4. Dezember 2025 vorgestellt und dort breit diskutiert. Hinweise aus dieser Sitzung wurden aufgenommen und haben noch einmal zu einer Überarbeitung geführt. Das Konzept wurde in der Sitzung des Steuerungsgremiums von genomDE am 16. Dezember 2025 verabschiedet.

2. Instrumente und Organisationseinheiten in einer sicheren Verarbeitungsumgebung

Der Begriff "Secure Processing Environment" ist definiert im Data Governance Act (EU 2022/868) als die physische oder virtuelle Umgebung und die organisatorischen Mittel, mit denen die Einhaltung der Anforderungen des Unionsrechts, wie der Verordnung (EU) 2016/679, insbesondere im Hinblick auf die Rechte der betroffenen Personen, der Rechte des geistigen Eigentums und der geschäftlichen und statistischen Vertraulichkeit, der Integrität und der Verfügbarkeit, sowie des geltenden Unionsrechts und des nationalen Rechts gewährleistet wird, und die es der Einrichtung, die die sichere Verarbeitungsumgebung bereitstellt, ermöglichen, alle Datenverarbeitungsvorgänge zu bestimmen und zu beaufsichtigen, darunter auch das Anzeigen, Speichern, Herunterladen und Exportieren von Daten und das Berechnen abgeleiteter Daten mithilfe von Rechenalgorithmen.

Die Datenverarbeitungsvorgänge werden durch einen Gesundheitsdatennutzer angestoßen. Die Anforderungen der Nutzer aus Versorgung und Forschung müssen mittels SPE angemessen umgesetzt werden können. Das Modellvorhaben Genomsequenzierung muss nachweisen können, dass es diese Art der Datennutzung angemessen unterstützt.

[1] Das vorliegende Konzept sollte auch in Zukunft regelmäßig fortgeschrieben und an die jeweils aktuellen technischen, rechtlichen und organisatorischen Entwicklungen angepasst werden.

[2] Dieser gehören an: Prof. Dr. Thomas Berlage (Fraunhofer FIT), Dr. Knut Kaulke (TMF), Prof. Dr. Oliver Kohlbacher (Universität Tübingen), Dr. Pascal Kraft (GHGA), Prof. Dr. Michael Krawczak (TMF - AG BioSysMed), Sebastian C. Semler (TMF), Prof. Dr. Oliver Stegle (GHGA) und Friedrich von Kessel (TMF).

Voraussetzung für eine Datennutzung ist die sichere Identifikation der natürlichen Personen, die die SPE nutzen (siehe auch BSI Technische Richtlinie TR-03107-1: Elektronische Identitäten und Vertrauensdienste im E-Government). Die dafür nötigen Prozesse und Maßnahmen werden hier nicht weiter betrachtet, da es sich um ein übergreifendes Problem für jegliche Gesundheitsdatennutzung handelt, dessen ad hoc Lösung im vorliegenden Kontext nur übergangsweise implementiert werden sollte. Es ist aber festzuhalten, dass ein Zugang mittels Heilberufeausweis und Telematik-Infrastruktur nicht ausreicht, da legitime Nutzer in der Forschung und Projektpartner im Ausland auf diese Weise nicht partizipieren könnten. Daher ist auch ein Identifikationsmechanismus aus dem Forschungssektor (z.B. aus den Bereichen GA4GH, Elixir, NFDI, EHDS) vorzusehen.

Im Rahmen des europäischen TEHDAS II-Projektes entsteht ein Dokument zu Spezifikationen für den Betrieb einer SPE. Dieses Dokument ist bereits in der aktuell vorliegenden Fassung nützlich, da es neben den konkreten Anforderungen der EHDS-Regulierung in weiten Bereichen andere Richtlinien für den Betrieb sicherer Informationssysteme (ISO27001/27002, NIS2) referenziert. Die operativen Anforderungen komplementieren das hier vorgestellte Konzept. Im Übrigen ist auch die BSI Richtlinie TR-03161 "Anforderungen an Anwendungen im Gesundheitswesen" zu beachten.

2.1 Prozess der Datennutzung

Basierend auf den FAIR-Prinzipien unterteilt sich ein Datennutzungsprozess in vier Schritte:

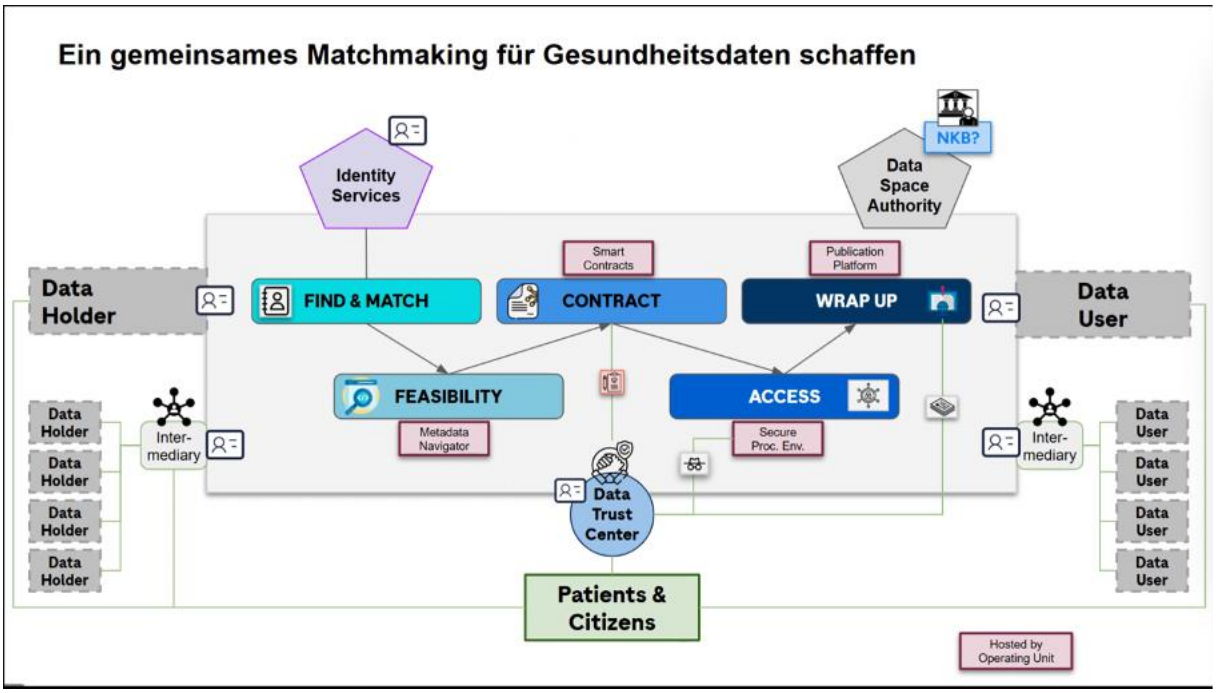
Auffindbarkeit (Findability)
Ein potenzieller Datennutzer prüft die Verfügbarkeit von geeigneten Daten in gewünschter Menge und Qualität. Dabei wird zunächst auf Metadaten zurückgegriffen, die von den Datenquellen publiziert und ggf. in Masterkatalogen aggregiert werden. Das Auffinden geht aber nahtlos in die Untersuchung der Machbarkeit ein, und je nach Anforderung müssen die Daten im Detail geprüft werden. Da ein Datenzugang erst auf Antrag gewährt wird, können in diesem Schritt nur anonyme Resultate geliefert werden. Allerdings ist es grundsätzlich möglich, eine detaillierte Anfrage an die Datenquellen zu stellen, die dort nur mit Zugang zu den Originaldaten beantwortet werden kann, als Antwort aber wiederum anonym ist (beispielsweise "Gibt es mindestens hundert Fälle mit Diagnose=X und Therapie=Y?").

Zugänglichkeit (Accessibility)
Die Bereitstellung setzt einen Antrag und eine Bewilligung (des access bodies) voraus. In dieser Bewilligung ist (unter anderem) der genaue Datenumfang festgelegt. Beitrag der Datenquelle ist es, genau die gewünschten Daten zusammenzustellen und der SPE einen Zugangskanal zu diesen anlassbezogenen Daten zu geben. Die Daten müssen nicht notwendigerweise übermittelt werden, aber es muss eine Programmierschnittstelle (API) geben, so dass grundsätzlich jedes freigegebene Datenelement für Verarbeitungsoperationen anwählbar ist (unabhängig davon, ob die Verarbeitung föderiert in der Datenquelle oder in einer separaten SPE-Instanz stattfindet).

Interoperabilität (Interoperability)
Der Datennutzer muss mit einer SPE interagieren können, die es erlaubt, simple oder komplexe Operationen auf Anforderung durchführen zu können. Dabei kann es sich um einfache Rechenbefehle (beispielsweise in einer Sprache wie Python oder R) handeln oder um komplexe Softwarebefehle (von Bioinformatik-Pipelines bis zu maschinellen Lernmodellen). Zwischenergebnisse bleiben in der SPE eingeschlossen, außer der Nutzerverbindung gibt es keine anderen Kommunikationskanäle nach außen (außer eventuell absolut notwendiger und gesicherter Verbindungen zu Terminologie- oder Referenzservern vertrauenswürdiger Institutionen).

Nachnutzbarkeit (Reusability)
Ergebnisse der Verarbeitung in der SPE müssen festgehalten und zur Nutzung übermittelt werden. Dabei ist (von Seiten des access bodies) zu prüfen, ob die durchgeführten Verarbeitungen und die auszuleitenden Ergebnisse zum genehmigten Nutzungszweck passen. Insbesondere ist zu prüfen, ob durch die Verarbeitung unzulässigerweise Personenbezüge hergestellt werden können.

Das vom BDI organisierte "Konvergenzboard Gesundheit" hat ein entsprechendes Prozessmodell entwickelt:



Auch kommerzielle Anbieter wie BC Platforms unterstützen einen entsprechenden Prozess.

Entscheidungen, wie die einzelnen Schritte dieses Modells ausgestaltet werden, wie die Aufgaben verteilt werden und inwieweit dies die verschiedenen Nutzungsszenarien umsetzbar abdeckt, sind Aufgabe der Governance (s. Abschnitt 5).

2.2 Architekturkomponenten

Damit die verschiedenen Schritte der Datennutzung so weit wie möglich automatisiert unterstützt werden können (unabhängig davon, wer die jeweiligen Funktionen festlegt, betreibt oder überwacht), sind eine Reihe von Funktionen erforderlich.

Ziel der Architektur ist es, die Hauptfunktionen unterschiedlichen Komponenten (Module) zuzuordnen, die möglichst einfache Schnittstellen und Kommunikationswege untereinander haben sollen. Gleichzeitig soll es möglich sein, für verschiedene Fälle auch unterschiedliche Varianten jeder Komponente zur alternativen Nutzung zu realisieren.

Grundlage einer verteilten Architektur sind Netzwerkknoten (hier nur kurz Knoten genannt), die miteinander über sichere Verbindungen kommunizieren können. Knoten können Daten halten (Datenknoten) und Verarbeitungsleistungen anbieten (Rechenknoten) oder beides. Im Modellvorhaben bilden die klinischen Datenknoten (KDK) und die Genomrechenzentren (GRZ) Datenknoten, die GRZ sind auch Rechenknoten. Weitere Rechenknoten können nach Bedarf hinzugefügt werden, zum Beispiel zur Implementation von Datendiensten. Alle hier beschriebenen Funktionen sind jeweils in einem oder mehreren Knoten lokalisiert.

Anhand der Schritte der Datennutzung gibt es die folgenden Komponenten:

1. Metadatenkatalog. Im Modellvorhaben hat jeder Datenknoten einen festgelegten Datenbestand. In Zukunft kommen aber weitere Datenbestände hinzu. Zur Identifikation und Unterscheidung und zur Beschreibung des Inhalts muss ein Datenknoten für jeden Datenbestand Metadaten bereitstel-

len. Da der Metadatenkatalog nur anonyme Daten enthält, kann er weitgehend öffentlich zugänglich sein. Eine verteilte Datenhaltung ist möglich, aber zur Erleichterung der Suche ist eine zentrale Aggregation (bzw. mindestens ein zentraler Zugang) vorteilhaft. Da vielleicht nicht alle Informationen im Metadatenkatalog selbsterklärend sind, sollte über den Katalog auch der Zugang zu einer persönlichen inhaltlichen Beratung möglich sein.

2. Machbarkeitsunterstützung. Ob Daten in Menge und Qualität für einen bestimmten Zweck geeignet sind, lässt sich oft nur durch Einsicht in die echten Daten herausfinden ("Wie viele Fälle gibt es mit Merkmal X, und gibt es genügend andere Fälle ohne dieses Merkmal?"). Es wäre eine große Erleichterung, wenn Machbarkeitsfragen bereits automatisiert beantwortet werden können, bevor ein formeller Antrag gestellt, bewilligt und umgesetzt ist. Erforderlich ist eine interaktive Umgebung zur Exploration (zum Beispiel in Form eines interaktiven Dashboards), bei der aber technisch-konzeptionell sichergestellt werden muss, dass sie nur aggregierte und nicht personenbezogene Daten darstellt.
3. Datenbereitstellung. Die für die Nutzung freigegebenen Daten sollten automatisiert anhand der Angaben im Antrag bereitgestellt werden. Durch die verteilte Datenhaltung würde sonst ein hoher, paralleler Aufwand in den Datenknoten für die manuelle Verarbeitung entstehen. Dafür sind eine automatisierte Datenselektion bei den Datenquellen, ein automatisierter Datentransport sowie eine automatisierte Aufbereitung in der SPE erforderlich. Entsprechende Funktionen finden sich in allen aktuellen Datenraumkonzepten. Im Rahmen der Bereitstellung werden auch die Zugangsrechte (signiert von der Datenzugangsstelle) durchgesetzt. Eine Protokollierung zur Nachvollziehbarkeit ist möglich. Eine sinnvolle Zusatzfunktion wäre das Einfrieren von Anfragen, so dass die Daten selbst nicht aufbewahrt werden müssen, eine exakte Rekonstruktion des zusammengeführten Datensatzes jedoch jederzeit möglich ist.
4. Nutzeridentifizierung.
5. Recheninstanz. In der SPE wird für jedes Projekt eine separate, von allen anderen Nutzern getrennte Instanz temporär bereitgestellt und nach Beendigung wieder gelöscht.
6. Ausführungsplattform. Eine SPE basiert auf einer Cloud-Infrastruktur (PaaS platform-as-a-service), die die IT-Ressourcen bereitstellt. Grundlage für den Betrieb dieser Plattform ist das SecDevOps Prinzip (Integration von Sicherheitsmaßnahmen und ständiger Software-Weiterentwicklung in den regulären IT-Betrieb).
7. Datendienste. Diese werden für vordefinierte und wiederkehrende Verarbeitungen eingerichtet. Ein Datendienst ist eine spezielle Softwareapplikation (SaaS software-as-a-service), die auf Basis der Ausführungsplattform und anderer Funktionen implementiert und durch den Plattformträger freigegeben wird.
8. Programmierumgebung (trusted research environment). Viele Forschungsvorhaben benötigen eine flexible Verarbeitung basierend auf spezialisierten Tools. Im Rahmen einer Programmierumgebung können Teams von natürlichen Personen Datenanalysen mithilfe der im Antrag angegebenen Softwarewerkzeuge durchführen.
9. Antragsportal. Damit beantragte Projekte automatisiert umgesetzt werden können, müssen Nutzungsanträge in einem Portal digital und strukturiert verwaltet werden. Aus den Anträgen leitet sich dann die Konfiguration der einzelnen Verarbeitungskomponenten ab (beispielsweise der Datenbereitstellung). Möglichst viele der dafür erforderlichen Informationen sollten direkt von den Antragstellern interaktiv eingegeben werden.
10. Infrastruktur Monitor. Alle Datennutzungen müssen protokolliert und bezüglich verdächtiger Vorkommnisse überwacht werden. Dies ermöglicht die Erkennung von unerwünschten Vorfällen (incidence detection, z.B. bei einer Infiltration des Systems) und deren schnelle Beseitigung (incidence recovery).
11. Software Repository. Dort werden alle verwendbaren Softwarekomponenten (Datendienste, Tools, Bibliotheken, ggf. abgesicherte Zugänge auf externe Informationsquellen) registriert, signiert und ggf. gespeichert.

2.3 Aufgabenverteilung

Grundlegende institutionelle Rollen in der Umsetzung der Dateninfrastruktur bekleiden der Plattformträger, die Vertrauensstelle und die Betreiber der Datenknoten.

Rolle des Plattformträgers
Als primäre Anlaufstelle für prospektive Nutzer des SPE trägt der Plattformträger die Verantwortung für den Betrieb des Antragsportals. Das Antragsportal übernimmt die zentrale Nutzerverwaltung sowie die Berechtigungsverwaltung gemäß der Datenbewilligung. Dies umfasst die Abbildung eines Authentifizierungs- und Autorisierungsprozesses sowie die Verwaltung der Rollenprofile und die Sicherstellung einer minimalen Rechtevergabe im Sinne des Minimierungsprinzips für jegliche Datennutzung.

Über den Betrieb der Antragsplattform hinaus liegt auch das Treffen rechtsverbindlicher Entscheidungen im Rahmen der Antragstellung beim Plattformträger. Im Kontext des Datenschutzes treten SPE-Betreiber nur als Auftragsdatenverarbeiter auf und müssen daher für alle Tätigkeiten beauftragt werden. Dem Plattformträger obliegt in diesem Rahmen die Auditierung der relevanten Prozesse (Antragsbewilligung, Projektlaufzeiten, Zugangskontrolle, etc.).

Um diese Entscheidungen treffen zu können, ist der Plattformträger darüber hinaus mit mehreren Stufen der Risiko-Analyse betraut. Dies umfasst zum einen die Risiko-Analyse der organisatorischen und technischen Rahmenbedingungen einzelner Anträge im Zuge des auditierten Antragsprozesses, erstreckt sich darüber hinaus aber auch auf die regelmäßige Risiko-Analyse des Systems als Ganzes (IT-Sicherheit der involvierten Komponenten, Dokumentation von Zwischenfällen, Risiken und Szenarien der Plattform als solcher sowie Mitigationsstrategien). Dies umfasst sowohl sog. Threat Modelling (die Identifikation von Risiken, bevor sie eintreten, als Teil einer strukturierten Analyse), die Etablierung von Schutzmaßnahmen gegen diese Risiken (Empfehlungen zu Architektur, Standards), als auch das Überwachen der Systeme auf Anzeichen eben jener Risiken. Wo der Plattformträger nicht selbst als Betreiber einer Plattform auftritt (und dies deshalb selbst tun kann), ist es seine Aufgabe, Empfehlungen auszusprechen und die Umsetzung zu auditieren, um den sicheren Betrieb der Infrastruktur zu gewährleisten.

Nachdem ein Antrag (nach Einreichung, Risiko-Analyse und positiver Entscheidung) bewilligt wurde und der Zugriff auf geschützte Daten im Rahmen des SPEs erfolgt, ist der Plattformträger verpflichtet, für die Zwecke des Datenschutzes und die Einhaltung technischer und juristischer Rahmenwerke ("Was kann der Nutzer mit den Daten tun? Was darf der Nutzer mit den Daten tun?") adäquate Kontroll- und Überwachungsprozesse zu etablieren. Es wird jedoch empfohlen, von einer kompletten technischen Unterbindung in Fällen nicht zulässiger Datennutzungen abzusehen, da die hierfür erforderlichen Vorkehrungen die Nutzung der SPE so stark einschränken könnten, dass ein Mehrwert für den Nutzer nicht mehr gegeben wäre. Stattdessen sollten juristische Rahmenwerke (Nutzungsbedingungen) erstellt, deren Einhaltung überwacht und Verstöße durch geeignete Maßnahmen gegenüber dem verantwortlichen Nutzer sanktioniert werden.

Über das Antragsportal hinaus übernimmt der Plattformträger die Verwaltung einer Software-Whitelist. Diese Liste dient als Grundlage eines Software-Repositoriums, das die SPE-Instanzen für den Bezug von Software nutzen. Diese Aufgabe hat zwei Kernkomponenten:
1. Die Entscheidungsautorität, welche Software für die Nutzung in der SPE freigegeben wird.
2. Den Betrieb einer Infrastruktur, die es den SPE-Instanzen erlaubt, diese Software zu beziehen.

Rolle der Vertrauensstelle
Da genomische Rohdaten ohne Informationsverlust nicht anonymisiert werden können, steht vor dem Zugriff auf diese Daten eine Pseudonymisierung. Die Vertrauensstelle erstellt und verwahrt nicht nur die primären Pseudonyme zur Speicherung in den Datenknoten, sondern auch projektbezogene Pseudonyme sowie deren Assoziation mit den originären Identitäten. Diese werden für den Zeitraum einer Datennutzung in einer SPE vergeben.

Hierfür ist eine technische Infrastruktur zu schaffen, die es den Datenknoten und SPE-Betreibern ermöglicht, eine Zuordnung von globalen Identitäten aller zu pseudonymisierender Entitäten auf projektbezogene Pseudonyme abzurufen. Zur Erreichung dieses Ziels ist eine hinreichend abgesicherte API zu betreiben, die von Datenknoten und SPE-Betreibern für den Abruf der Pseudonymtabellen genutzt werden kann. Identitäten in den bereitzustellenden Daten werden durch deren projektbezogene Pseudonyme ausgetauscht, bevor die Daten herausgegeben werden.

Darüber hinaus kann es projektbezogen erlaubt werden, die Pseudonymisierung an den SPE-Betreiber zu delegieren. Bei besonders umfangreichen Datensätzen, die in mehreren Projekten genutzt werden, ermöglicht dies dem SPE-Betreiber, multiple Pseudonyme des gleichen Genomdatensatzes aufzulösen, um eine Duplizierung der Daten zu vermeiden.

Rolle der Datenknoten

Die Datenknoten übernehmen sowohl für die klinischen als auch für die genomischen Daten primär deren Bereitstellung sowie die Bereitstellung der zugehörigen Metadaten. Es sind entsprechende Vorkehrungen zu treffen, um diese Aufgabe auch in Zukunft zuverlässig erfüllen zu können. Um ihre Rolle ausüben zu können, müssen die Datenknoten vom Plattformträger über eventuelle Freigaben informiert werden, da ohne die Beauftragung durch den Plattformträger eine Weitergabe von Daten nicht erfolgen darf. Wichtig ist hierbei der Aspekt, dass jede Zustimmung zur Forschungsnutzung von Daten nicht notwendigerweise dauerhaft ist – es kann passieren, dass im Laufe eines SPE-Projekts diese Zustimmung von einem oder mehreren Patienten zurückgezogen wird. Die Datenknoten werden darüber gegebenenfalls vom Plattformträger informiert und müssen sicherstellen, dass die Daten danach nicht mehr herausgegeben werden. Dieselbe Information ist auch an den SPE-Betreiber zu geben, der die betroffenen Daten aus dem SPE entfernen oder unbrauchbar machen muss.

Ergebnisse einer Verarbeitung in der SPE sollten wiederverwertbar sein. Dafür müssen Datenknoten in die Lage versetzt werden, Ergebnisse in ihren Datenbestand zurückzuspielen, z.B. um eine erneute Berechnung für künftige Projekte zu vermeiden. Die Entscheidung, welche Daten gespeichert werden sollen, obliegt dem Plattformträger, das Rückspielen übernimmt die SPE-Instanz, aber der Betrieb einer Infrastruktur für das Rückspielen obliegt den Datenknoten. Werden zum Beispiel im Rahmen eines Projekts die Rohdaten gegen ein anderes Referenzgenom aligniert, kann der Plattformträger entscheiden, dass der hierfür erforderliche hohe Aufwand nicht wiederholt werden sollte und deshalb die Daten aus dem Projekt dem ursprünglichen Datensatz hinzugefügt werden können.

3. Konzept und technischer Demonstrator SPE genomDE

Im Rahmen eines technischen Demonstrators sollen die grundlegenden Herausforderungen einer SPE in vereinfachtem Umfang durchgespielt werden. Das bedeutet, dass all diesen Herausforderungen qualitativ, aber nicht quantitativ begegnet wird. Die daraus resultierende Vereinfachung des Szenarios erlaubt das schnelle Sammeln von realen Nutzungsdaten und Erkenntnissen aus dem Betrieb der erforderlichen Infrastruktur für die bessere Planung und Umsetzung eines entsprechenden Projekts in der Zukunft. Im Fall des technischen Demonstrators wird die Frage der Antragsbewilligung bewusst ausgeklammert – es wird davon ausgegangen, dass ein Projekt oder mehrere Projekte vorliegen, zu denen der Plattformträger eine Freigabe erteilt hat. Darauffolgend unterteilt sich der Prozess auf der Seite eines SPE-Betreibers in drei Phasen.

3.1 Phasen der SPE-Bereitstellung

Phase 1: Gesteuerte Datenbereitstellung und Initialisierung der SPE (Access)

Diese Phase wird durch eine extern erteilte, digital signierte Nutzungsgenehmigung angestoßen. Sie umfasst die Einrichtung der Arbeitsumgebung und die sichere Anbindung der genehmigten Daten.

1. Instanziierung der SPE:
   - Basierend auf der Genehmigung wird automatisiert eine dedizierte und isolierte SPE-Instanz (Research Analytics Zone (RAZ) oder Analytikzone) für das spezifische Nutzungsvorhaben erstellt. Alternativ erfolgt eine anwendergebundene Provisionierung des Datenzugangs.
   - Diese Instanz wird mit den exakt genehmigten Software-Werkzeugen aus dem zentralen Software Repository und den definierten Rechen- und Speicherkapazitäten konfiguriert. Die Nutzung der Ressourcen erfolgt isoliert, die Ressourcen selbst müssen jedoch nicht isoliert sein. (Beispielsweise sind die Compute Nodes in einem Cluster nicht für ein einzelnes Projekt reserviert, sondern können nacheinander verschiedene Projekte abarbeiten. Dabei ist aber darauf zu achten, dass Reste des vorherigen Projektes nicht mehr zugänglich sind.)

2. Autorisierter Datenzugriff:
   - Die SPE-Instanz nutzt die Genehmigung, um über ihre Abfragezone (auch Query Management Zone (QMZ)) eine gesicherte Verbindung zu den Datenzonen (auch Secure Data Zone (SDZ)) der relevanten Datenquellen (KDK, GRZ) aufzubauen.
   - Nach dem "Code-to-Data"-Prinzip werden die Daten nicht direkt dem Nutzer, sondern ausschließlich der kontrollierten SPE bereitgestellt. Dies geschieht entweder durch
     - Föderierte Analyse: Die Analyse-Skripte werden in der SPE ausgeführt, aber die Abfragen werden zur Verarbeitung an die Datenquellen gesendet. Nur die aggregierten Ergebnisse kehren in die SPE zurück.
     - Anlassbezogene Datenübertragung: Eine temporäre Kopie der genehmigten, pseudonymisierten Daten wird verschlüsselt in den Speicher der SPE-Instanz geladen. Wie in 2.3 beschrieben, kann an dieser Stelle die Pseudonymisierung auch an den SPE-Betreiber delegiert werden, falls dieselben Daten in mehreren Projekten innerhalb desselben SPE zur Verfügung gestellt werden. Dies vermeidet die Duplizierung der Daten.
   - Der Datennutzer hat zu keinem Zeitpunkt direkten Kontakt zu den angebundenen Quellsystemen oder eine Kontrolle über den Datenübertragungsprozess. Alle Zugriffe werden protokolliert.

<u>Phase 2: Sichere Datenverarbeitung in der gekapselten Umgebung (Interoperate)</u>

Dies ist die Kernphase, in der der Datennutzer seine Analysen innerhalb der geschützten "Airlock"-Umgebung durchführt.

1. Durchführung der Analyse:
   - Der autorisierte Datennutzer erhält einen gesicherten Zugang (Login) zu seiner persönlichen SPE-Instanz.
   - Innerhalb dieser Umgebung kann er die bereitgestellten Daten mit den genehmigten Werkzeugen (z. B. R, Python, Bioinformatik-Pipelines) analysieren.
   - Jegliche Netzwerkkommunikation aus der SPE nach außen ist blockiert, um einen unkontrollierten Datenabfluss zu verhindern. Als Ausnahmen hierzu können von Seiten des SPE-Betreibers sogenannte Allow Lists eingerichtet werden. Hierzu erfragt zunächst der SPE-Betreiber beim Plattformträger die Vertrauenswürdigkeit einer Datenquelle. Ist eine Quelle als vertrauenswürdig eingestuft, kann der SPE-Betreiber von einer Blockade der Netzwerkkommunikation mit den entsprechenden Endpunkten absehen. Diese Entscheidung und Umsetzung ist dem Audit-Trail der SPE-Instanz hinzuzufügen.
   - Alle Aktionen des Nutzers und alle Rechenprozesse werden durch den Infrastruktur Monitor lückenlos protokolliert.

2. Verwaltung von Zwischenergebnissen:
   - Alle während der Analyse erzeugten Skripte, Modelle und Zwischenergebnisse verbleiben zwingend innerhalb der SPE. Sie werden im temporären Speicher der Instanz sicher abgelegt und können für die weitere Bearbeitung wiederverwendet werden, dürfen die Umgebung aber nicht verlassen.

Phase 3: Geprüfte Ergebnisausleitung und Beendigung (Reuse)

Diese Phase regelt den kontrollierten Export der finalen Ergebnisse und die ordnungsgemäße Stilllegung der Umgebung.

- Beantragung der Ergebnisausleitung (Output Control):
  - Der Datennutzer stellt aus der SPE heraus einen Antrag auf Export seiner finalen, aggregierten Ergebnisse (z. B. Statistiken, Grafiken, Tabellen).
  - Eventuell auf Basis des Zugriffsantrags nicht erforderlich (Antragsportal).
- Prüfung auf Re-Identifizierungsrisiko:
  - Die zur Ausleitung beantragten Ergebnisse werden einem automatisierten und/oder manuellen Prüfprozess unterzogen. Dabei wird sichergestellt, dass sie dem genehmigten Nutzungszweck entsprechen und keine direkten oder indirekten personenbeziehbaren Informationen enthalten (soweit nicht zweckerforderlich).
- Freigabe und Übermittlung:
  - Nur nach erfolgreicher Prüfung werden die Ergebnisse aus der SPE ausgeleitet und dem Nutzer über einen sicheren Kanal zur Verfügung gestellt. Für diesen Zweck stellt der SPE-Betreiber die notwendige Infrastruktur bereit. Optionen für eine solche Übergabe der Daten sind SFTP oder ein Download-Portal mit https, das dieselbe Authentifizierung und Autorisierung nutzt, wie das Antragsportal (siehe 2.3.).
- De-Provisionierung der SPE:
  - Archivierung der generierten Ergebnisse.
  - Nach Abschluss des Projekts oder nach Ablauf der genehmigten Nutzungsdauer wird die gesamte SPE-Instanz inklusive aller temporären Daten und Zwischenergebnisse sicher und unwiderruflich gelöscht. Dieser Vorgang wird protokolliert und beendet den Lebenszyklus der Verarbeitung.

Bei hochgradig individuellen Forschungsprojekten wie zum Beispiel dem Use-Case "Forschungsumgebung R / Python" aus Abschnitt 4.1 ist nicht davon auszugehen, dass generierte Ergebnisse sehr nützlich für andere Forschende oder sehr teuer zu generieren sind, und somit eine erneute Berechnung vermieden werden sollte. Für Dashboards (mit einer inhärenten Ontologie der Daten) ist die Sachlage jedoch grundlegend anders: Die Repräsentation eines Patienten (anhand der Daten in den klinischen Datenknoten oder der genomischen Daten) in den Datenstrukturen eines Daten-Dashboards muss nicht mehrfach berechnet werden und sollte deshalb in Phase 3 eines Projekts in die entsprechende Datenquelle zurückgespielt werden. In diesem Fall werden in einer SPE für jeden Patienten die für das Dashboard erforderlichen Daten berechnet, anonymisiert und dann als Ausgabe zur Verfügung gestellt. Darüber hinaus kann dieser Datensatz in den jeweiligen Datenknoten vorgehalten werden, falls er in der Zukunft erneut genutzt werden soll. Dasselbe gilt für den Use Case “patients-like-mine”. Dort könnte beispielsweise ein Index aller Phänotypen vorgehalten werden, damit dieser nicht für jede Anfrage erneut berechnet werden muss.

Für solche Projekte kann gegebenenfalls ein langlebiges SPE etabliert werden, das kein vorgesehenes Enddatum hat und die aggregierten Daten zentral zwischenspeichert. Für diese Option müssten besondere Risikoanalysen erstellt und strengere Sicherheitsvorkehrungen etabliert und umgesetzt werden.

3.2 Technisches Konzept für die Skalierung von SPE für die Sekundärnutzung

Für viele Use-Cases – vor allem solche, die nicht genomische Rohdaten, sondern vorverarbeitete Daten nutzen – ist keine gesonderte Betrachtung der technischen Machbarkeit erforderlich. Wenn nur vorverarbeitete Daten genutzt werden, können diese in einer SPE-Instanz zusammengeführt und dort verarbeitet werden.

Projekte, die genomische Rohdaten nutzen wollen, sind hier als eingeschränkt zu betrachten. In diesem Fall kann es (je nach Größe des Projekts) zu aufwändig sein, die Verarbeitung in einer einzigen SPE

durchzuführen. In diesem Fall ist von den Antragstellern eine Unterteilung des Forschungsvorhabens in einen föderierbaren und einen nicht föderierbaren Anteil sowie eine Spezifikation des Datenaustausches zwischen den beiden Anteilen zu gewährleisten. Die dem Forschenden zur Verfügung gestellte SPE-Instanz ist im Fall einer erteilten Freigabe des Projekts lediglich für den nicht-föderierten Anteil konzipiert und der föderierte Anteil ist von den Datenknoten mit SPE-Betreibern in Kooperation zu gewährleisten. Die Freigabe dieser Auftragsdatenverarbeitung ist Teil der ursprünglichen Projektfreigabe.

Es ist durch den Plattformträger zu prüfen, ob alle KDK ressourcentechnisch (Stichwort: Vorhaltekosten!) in die Lage versetzt werden sollten, eine föderierte Verarbeitung zu ermöglichen. Ist dies nicht machbar, könnten sich bestimmte Forschungsvorhaben als prohibitiv aufwändig erweisen. Dies gilt jedoch grundsätzlich für alle avancierten Formen der Datenverarbeitung, beispielsweise mit aufwendigen KI-Modellen. In solchen Fällen gilt, dass die erforderliche Infrastruktur für die Verarbeitung im Rahmen der Planung und Finanzierung des Forschungsvorhabens geschaffen werden muss.

Diese Struktur ermöglicht es den Datenknoten, alle erforderlichen Ressourcen zu nutzen, um einen Verarbeitungsauftrag zu erfüllen. Wenn diese Vorarbeit erfolgt ist, kann die Ausleitung der Ergebnisdaten gemäß der dafür zur Verfügung gestellten Spezifikation des Datenaustausches erfolgen. Die Ergebnisdaten werden dann von den Datenknoten wie in einem nicht-föderierten Projekt in eine SPE-Instanz eingespielt.

Beispiel: Wenn in einem Forschungsvorhaben alle Genome mit einem sog. Custom Variant Caller reanalysiert werden sollen, kann dies nicht in einer einzigen SPE-Instanz erfolgen. Dafür müsste der komplette Datensatz in eine einzige Infrastruktur kopiert werden (was weder technisch noch aus Sicherheitsgründen sinnvoll ist). Hierfür würden die Forschenden ein Skript zur Verfügung stellen, welches auf jedem einzelnen Genom ausgeführt werden soll. Jeder beteiligte Datenknoten kann die Ausführung dieser Berechnung für die von diesem Datenknoten gehaltenen Daten an einen SPE-Partner übergeben, der die Berechnungen optimal durchführen kann (zum Beispiel, weil der SPE-Partner eine besonders schnelle Verbindung zu dem Datenspeicher hat oder aufgrund seiner Erfahrung oder verfügbaren Infrastruktur besonders geeignet ist). Hierfür kann es erforderlich sein, sekundäre Daten von anderen Datenknoten zu beziehen (etwa klinische Daten zu den jeweiligen Patienten). Nach Abschluss der Berechnung werden die Ausgabedateien dieses ersten Schritts bei jedem Datenknoten gesammelt und in die SPE-Instanz für die Forschenden eingespielt. Dies entspricht einer Map-Reduce Implementierung von föderierter Datenverarbeitung. In solchen Fällen wird vom Antragsteller eine Beurteilung der Föderierbarkeit der Ausführung als Teil des Antrags erwartet.

Wenn ein föderiertes Projekt abgeschlossen ist, ist ein Bericht über die Umsetzung des Projekts zu erstellen. Dieser soll erfassen, wie die Umsetzung verlaufen ist und welche Erkenntnisse dabei gewonnen wurden. Mit dem Ziel der effizienten Durchführung föderierter Datenverarbeitung ist zu erfassen, welche Probleme in diesem Prozess vorliegen und welche Infrastrukturen sich als geeignet erwiesen haben, solche Analysen durchzuführen. Ziel dieser Arbeit ist es, eine Automatisierung des Prozesses zu planen und umzusetzen.

3.3 Record-Linkage Genomdaten und klinische Daten

Für nahezu jede Verarbeitung von Rohdaten in den GRZ ist eine Verknüpfung mit den klinischen Daten aus den KDK erforderlich, um eine passende Annotation der Genomdaten bezüglich klinischer Merkmale zu erzielen und Datensätze mit bestimmten Merkmalen zu selektieren. Diese Verknüpfung muss über die Vertrauensstelle erfolgen. Nur die Vertrauensstelle weiß, in welchem KDK die zugeordneten Daten zu finden sind. Es ist ein Protokoll erforderlich, das die gewünschten Pseudonyme im Block an die verschiedenen Knoten verteilt, diesen auch noch ein temporäres Pseudonym für die Abfrage zur Verfügung stellt, damit sie die erforderlichen Daten re-pseudonymisiert an die SPE liefern können. Die SPE sollte nicht in der Lage sein, die Assoziation der Ursprungspseudonyme zu erschließen. Ein mit dem RKI diskutiertes Protokoll ist als Anhang beigefügt.

4. Use Cases der Datennutzung

Eine SPE im Kontext von genomDE muss die Anforderungen der Nutzergruppen in der lernenden Versorgung niederschwellig erfüllen können. Zu hohe Hürden für die Beantragung und Benutzung sowie zu hohe Aufwände (technisch und organisatorisch) würden die Zwecke der Datenerhebung konterkarieren.

4.1 Modellierung von Use Cases

Um das erstellte Konzept zu evaluieren, sollen essenzielle Use Cases so weit wie möglich modelliert werden. Diese Use Cases sollen möglichst allgemein und repräsentativ formuliert werden, aber so konkret, dass die Umsetzung im erstellten Konzept bewertet werden kann.

Im Rahmen von drei Workshops wurden vier Szenarien erarbeitet.

UC1	**UC2**	**UC3**	**UC4**
Datenübersicht (Dashboard)	**Ähnlichkeitssuche ("Patients-like-mine")**	**Qualitätssicherung + Forschung mit Genomdaten (GRZ)**	**Adaptierbare Umgebung (R, Python)**
Ein vorprogrammiertes Dashboard ermöglicht interaktiv eine Übersicht über Fallzahlen, Merkmalsausprägungen und selektive Kohorten	Klinische Nutzer können nach detaillierten Merkmalsausprägungen suchen und bekommen pseudonymisierte Einzelprofile	Reprozessierung der Genomdaten komplett oder teilweise aufgrund gesuchter Merkmale oder zur Weiterentwicklung der Bioinformatik	Allgemeine Forschungsvorhaben mit Datenzusammenführung mit Krebsregisterdaten oder Kassendaten
Nützlich für das Modellvorhaben und Entlastung des Plattformträgers	Essenziell für das Modellvorhaben und Unterstützung der Machbarkeitsanalyse	Essenziell für das Modellvorhaben und Kollaboration EU-Konsortien	Essenziell für die Forschung und den EHDS
Zugang (anonyme Daten) weitgehend möglich	Zugang für Personen mit relevantem Versorgungsauftrag (auch außerh. MV)	Zugang nur für Bioinformatik-Spezialisten	Zugang auf Antrag
Ausprägungen: Qualitätssicherung, Evaluierung, Machbarkeit	Ausprägungen: Seltene Erkrankungen, Onkologie, Studienrekrutierung	Ausprägungen: Reprozessierung, Sonderfälle, Variantenstatistik	
Klinische Datenknoten	Klinische Datenknoten + Genom-Rechenzentren	Genom-Rechenzentren + Klinische Datenknoten	Eigener SPE-Knoten

4.2 Funktionalität der Use Cases

a) UC1 Datenübersicht

Es gibt eine große Bandbreite von Möglichkeiten, einen kompletten Datensatz in seinen Eigenschaften zu visualisieren, ohne dabei personenbeziehbare Details zu offenbaren. Dies beginnt bei einfachen Statistiken über die verschiedenen Untergruppen (SE vs Onko, Anzahl Follow Ups, etc). Diese können als Monitoring auch auf die Leistungserbringer heruntergebrochen werden (Anzahl der Fälle, Erfolgsstatistiken - aus Wettbewerbsgründen eingeschränkter Nutzerkreis). Zur Datenqualitätssicherung kann angezeigt werden, wie sich die Merkmalsausprägungen verteilen (zum Beispiel Nutzung der verschiedenen Diagnose-Codes oder HPO-Terms). Eine solche Übersicht kann auch eine erste Hilfe bezüglich der Mach-

barkeit von Forschungsprojekten sein, indem angezeigt wird, wie viele Fälle mit bestimmten Eigenschaften in der Datenbasis vorhanden sind. Schließlich ist anzunehmen, dass zielgerichtete Übersichten auch die Evaluation des Modellvorhabens teilweise oder sogar ganz unterstützen können.

Technisch gesehen lassen sich solche Datenübersichten als Web-Anwendungen, z.B. in Form von Dashboards oder Live Monitoring, mit Hilfe von kommerziellen und Open-Source Werkzeugen realisieren. Diese Werkzeuge erwarten die zugrundeliegenden Daten typischerweise in einer Datenbank, die den Werkzeugen lokal zur Verfügung steht. Man sollte aber in vielen Fällen auch eine reduzierte Datenbasis (als sogenannte materialisierte Sicht) durch eine föderierte Abfrage aller relevanten Datenknoten erstellen können. In den meisten Fällen ist der Programmieraufwand mit den entsprechenden Werkzeugen vergleichsweise gering, so dass problemlos verschiedene separate Datenübersichten für unterschiedliche Zwecke und Nutzergruppen erstellt werden können.

b) UC2 Ähnlichkeitssuche

Der Ansatz "patients-like-mine" basiert darauf, zugeschnitten auf einen bestimmten Fall eine Kohorte ähnlicher Fälle zu identifizieren und die Eigenschaften dieser Kohorte als Entscheidungsunterstützung zu nutzen[3]. Dabei sollten sowohl die zu verwendenden Kriterien für die Ähnlichkeit einstellbar sein als auch die anzuzeigenden Eigenschaften der Kohorte (andernfalls wäre es ein automatisiertes Medizinprodukt).

Auch wenn dies technisch einer Datenübersicht ähnelt, liegt hier die Erwartung zugrunde, dass die relevante Kohorte auch sehr klein werden kann, bis hin zum Einzeldatensatz (falls nötig, beispielsweise bei seltenen Erkrankungen). Außerdem erfordert die geeignete Auswahl von Kriterien eine hohe medizinisch-wissenschaftliche Kompetenz. Typische Nutzergruppen sind Expertinnen und Experten in der lernenden Versorgung, in der Onkologie bei der Vorbereitung eines molekularen Tumorboards, allgemein in der Unterstützung der Rekrutierung für klinische Studien und zur Unterstützung der zielgerichteten Diagnose bei seltenen Erkrankungen (Identifizierung möglicher genomischer Varianten aus vergleichbaren Fällen). Im letzteren Fall steht am Ende möglicherweise (in der Versorgung) die Anforderung einer Fallidentifizierung, um mit der behandelnden Stelle eines anderen Patienten Kontakt aufzunehmen und weitere Details der Fälle abzugleichen.

Da die Ähnlichkeitssuche potenziell tiefer und detaillierter in die Daten schaut, kann sie auch eine Unterstützung in der zweiten Phase einer Machbarkeitsuntersuchung spielen. In diesem Fall ist natürlich eine Fallidentifizierung ausgeschlossen und möglicherweise wird die Größe der Kohorte nach unten beschränkt, um die Funktion ohne größeren Prüfaufwand verfügbar machen zu können.

Es ist davon auszugehen, dass es mehrere Varianten der Ähnlichkeitssuche geben wird (Seltene Erkrankungen, Onkologie, weitere). Da sie alle auf das gleiche Datenmodell und ähnliche Merkmale zugreifen, können sie vermutlich auf der gleichen technischen Basis realisiert werden. Es gäbe also verschiedene interaktive Module, die sich auf die gleiche Art der Datenbereitstellung stützen. Die Datenbereitstellung bzw. Suche sollte föderiert aus den Datenknoten erfolgen, um die Aggregation eines umfangreichen Datensatzes mit zahlreichen Merkmalen zu vermeiden.

c) UC3 Qualitätssicherung und Forschung mit Genomdaten

Dieser Use Case behandelt alle Vorhaben, die genomische Rohdaten benötigen und daher aus Ressourcengründen weitgehend in den mit den GRZ assoziierten SPEs ausgeführt werden müssen.

In der Bioinformatik geschieht die Verarbeitung genomischer Daten in der Regel mit einer Kette von parametrisierten Werkzeugen. Eine typische Pipeline zur Verarbeitung von Rohdaten enthält unter anderem Schritte zum Mapping der Teilsequenzen auf ein Referenzgenom und die Extraktion von Varianten

[3] Gombar, S., Callahan, A., Califf, R. et al. It is time to learn from patients like mine. npj Digit. Med. 2, 16 (2019). https://doi.org/10.1038/s41746-019-0091-3

zum Referenzgenom (Variant Calling). Ein Werkzeug greift jeweils auf die Ergebnisse des vorigen Schrittes zu. Die Ergebnisse werden jeweils als Dateien im lokalen Speicher abgelegt.

Für verschiedene Zwecke werden auch verschiedene Pipelines verwendet (unterschiedliche Algorithmen und Parameter). Die Bioinformatik entwickelt sich ständig weiter. Im Zuge dessen werden sich die im Modellvorhaben verwendeten Pipelines ändern, zudem verwenden unterschiedliche Teilnehmer möglicherweise unterschiedliche Pipelines.

In genomDE müssen daher verschiedene Pipelines parallel und nacheinander auf die originalen Sequenzdaten angewendet werden können. Üblicherweise speisen sich die Pipelines aus Open Source Tools der internationalen Community sowie aus wenigen kommerziellen Anbietern. Die Daten liegen lokal in den GRZ vor, eine Zusammenführung ist aus Volumengründen nur in Ausnahmefällen möglich, daher werden zentrenübergreifende Analysen parallel föderiert ausgeführt.

Für Zwecke der Qualitätssicherung benötigen die Analysen Zugriff auf ausgewählte klinische Daten, beispielsweise die Art der Erkrankung, Hauptdiagnose u.ä., um nur für eine bestimmte Studie relevante Daten zu analysieren.

Ergebnisse der Pipelines können selbst wieder neue Datenquellen darstellen. Beispielsweise könnten Statistiken über genomische Merkmale und Merkmalskombinationen erstellt werden, die dann über alle GRZ hinweg in eine eigene Datenquelle aggregiert werden können.

d) UC4 Adaptierbare Umgebung

Eine adaptierbare Umgebung mit Programmiermöglichkeit entspricht in den Anforderungen weitgehend den in verschiedenen Kontexten, insbesondere im EHDS vorgesehenen sicheren Ausführungsumgebungen bzw. trusted research environments (TRE). In der Regel wird davon ausgegangen, dass Daten anlassbezogen zusammengeführt, ggf. über die Vertrauensstelle verknüpft und lokal verarbeitet werden können.

Eingebrachte Codes und Werkzeuge müssen antragsbezogen geprüft werden. Verschiedene Communities werden sich vermutlich auf gewisse Sätze von Standardwerkzeugen einigen, die dann als geprüfte Umgebung zur Verfügung gestellt werden. Solche Umgebungen werden auch unterschiedliche Anforderungen an Speicher- und Rechenkapazität haben.

4.3 Konzeption

Das DARE UK Konsortium hat ein Architekturmodell für einen Gesundheitsdatenraum mit sicherer Verarbeitung[4] vorgestellt. Es basiert auf einer Untersuchung bisheriger Umsetzungen sowie der Architekturen in GAIA-X, SIMPl, IDSA und X-Road. Es deckt den gleichen Anwendungsbereich ab wie genomDE. Die Konzepte lassen sich zudem gut auf genomDE abbilden und bilden einen direkten Bezugspunkt für die SPE-Konzeption in genomDE.

Die DARE UK Federated Architecture trifft die grundsätzliche Entscheidung, alle Knoten eines Netzwerkes potenziell mit Datenspeicherung und Datenverarbeitung auszustatten. So können in einem einheitlichen Rahmen verschiedene Formen der verteilten Verarbeitung, von der anlassbezogenen Zusammenführung bis zur vollständig föderierten Analyse abgebildet werden. Alle Datenknoten können optional mit einer sicheren Verarbeitungsmöglichkeit ausgestattet werden.

[4] Federated Architecture Blueprint V 2.2, 2024

Für einen solchen Knoten unterscheidet die Architektur drei getrennte Zonen, die einzeln auftreten oder in einem Knoten miteinander kombiniert werden können.

- Die Secure Data Zone (SDZ, *Datenzone*) speichert Gesundheitsdaten. Diese Daten können von den anderen Zonen lokal benutzt werden oder an andere Datenknoten in deren SDZ übermittelt werden. In genomDE entspricht dies den KDK bzw. GRZ bezogen auf ihre Datenspeicherung.
- Die Query Management Zone (QMZ, *Abfragezone*) ermöglicht die Ausführung von Algorithmen, insbesondere für die automatisierte Datenbereitstellung. Mit "Query" ist hier eine Abfrage gemeint, die in klassischen Datenbanksprachen formuliert werden kann. Ebenso kann eine solche Abfrage aber einen Satz von Softwareartefakten (Docker, Codemodule, Konfigurationsfiles, etc.) spezifizieren, die auf den Daten ausgeführt werden sollen. Ersteres passt eher zu den KDK, letzteres ist unmittelbar auf die GRZ anwendbar. Man kann diesen Ansatz als "low code" interpretieren, da die Gesamtabfrage variabel ist, sich aber auf vordefinierte/vorgeprüfte Werkzeuge stützen muss.
- Die Research Analytics Zone (RAZ, *Analytikzone*). Diese Zone unterstützt die Aktivitäten von Forschern oder Anfragenden und ermöglicht ihnen Artefakte und Zwischenresultate (temporäre Daten) des Projektes in einer sicheren Umgebung zu verwalten. Diese Zone ist bisher in genomDE nicht ausgestaltet und enthält Funktionen, wie sie etwa für die SPE des FDZ geplant sind.

Dieses Zonenmodell ermöglicht eine modulare Beschreibung und Implementierung unterschiedlicher Szenarien. Beispielsweise kann es isolierte SDZ geben. Dies entspricht in genomDE einer Klinik, die an ein klinisches Netzwerk angeschlossen ist. Daten werden lokal gespeichert, aber auch über das Netzwerk an den KDK weitergeleitet.

Die QMZ können aber auch auf die KDK verteilt werden. Der zentrale Knoten verteilt dann die Abfragen und aggregiert die Ergebnisse (föderierte Verarbeitung). In beiden Fällen würden die QMZ auf Basis vorgeprüfter Softwareartefakte operieren. Die Softwareartefakte stehen dabei unter der Governance des Plattformträgers und werden von einem oder mehreren Servern ausgeliefert, die Teil der Infrastruktur sind (Sicherheit gegen die Manipulation von Softwaremodulen).

Ein Knoten, der einen Datendienst betreibt, umfasst immer eine RAZ, in die sich natürliche Personen als Nutzer einloggen können. Hier können einfach zu bedienende interaktive Werkzeuge (Dashboards, Drill-Down-Tools, etc.) zur Verfügung stehen, um häufige Use Cases zu adressieren. Diese Tools können dann je nach Use Case auf lokale Daten zugreifen, oder lokale oder verteilte Queries benutzen. Der Datendienst würde ebenso unter die Governance der Architektur fallen.

Schließlich kann ein Datendienst in einer RAZ auch eine komplette Programmierumgebung mit R und Python und den benötigten Community Tools umfassen. DARE UK fasst das unter dem Begriff "Projekt" zusammen, also eine temporäre Struktur nach Antrag, die Zugriff durch ein Team von Anwendern erlaubt.

Zwischen den einzelnen Zonen und zwischen den Teilnehmern des Datenraums können aufgrund dieser Architektur die Schnittstellen weitgehend standardisiert werden.

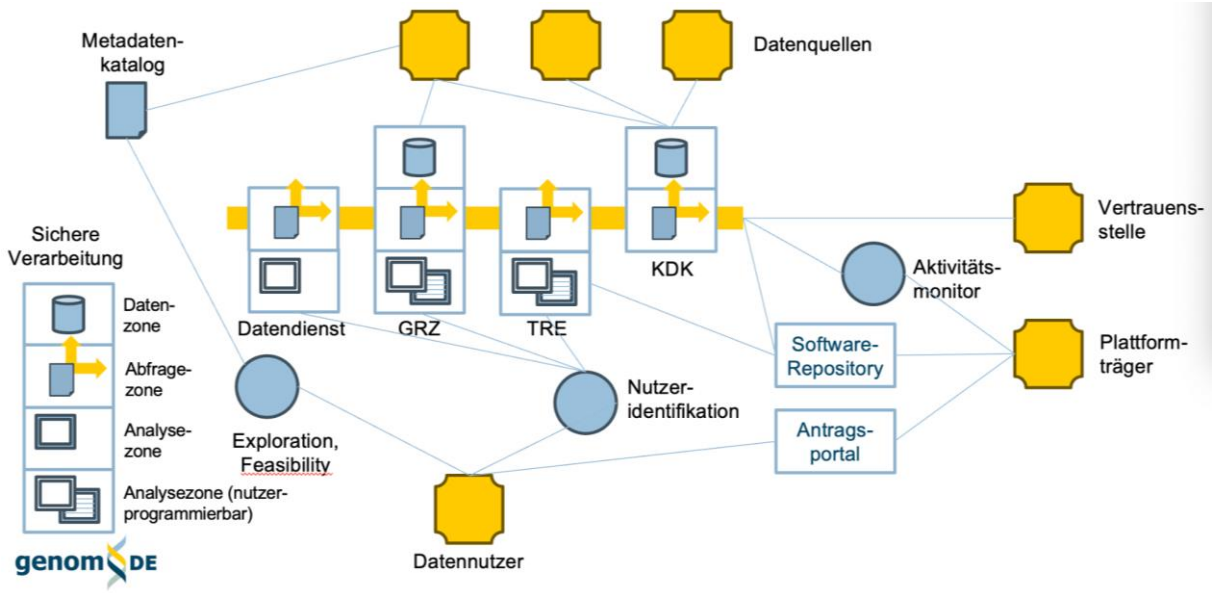
Die Architektur unterstützt eine strukturierte Governance.

Auf die Datenzone haben nur die sogenannten Data Custodians der jeweiligen datenschutzrechtlichen Controller Zugriff. Sie können bei Bedarf manuelle Transfers initiieren und kontrollieren außerdem die Datenzugriffsrechte anderer Komponenten/Projekte.

Die Abfragezone ermöglicht einen geschützten Raum für Softwareartefakte sowie für verteilte Verarbeitung und die automatisierte Datenbereitstellung. Auch hier haben Datennutzer keinen Zugriff, sondern nur die autorisierten Vertreter der jeweiligen Governance.

Die Analytikzone ist der Zugang zu einer SPE und steht ebenfalls unter dem Monitoring des Governance-Bodies. Hier gibt es unmittelbare Anschlüsse an den Nutzungsantrag und Prüfungen der Durchführung eines daraus ggf. resultierenden Projekts. Zulassungen von Nutzern sind ebenso wie Zulassungen von interaktiven Werkzeugen hier zu regeln.

In der folgenden Grafik ist ein Architekturmodell für genomDE gezeigt, das die Aspekte von DARE UK integriert:



Im Kern werden vier verschiedene Arten von Knoten unterschieden:

- Ein KDK enthält eine Datenzone für die lokalen Daten sowie eine Abfragezone für die Verarbeitung im Auftrag anderer Knoten.
- Ein GRZ verfügt zusätzlich über eine Analytikzone, in der die typischen Bioinformatikwerkzeuge für Nutzer verfügbar gemacht werden.
- Ein spezifischer Datendienst ist ein Knoten ohne eigene Daten, der eine Abfragezone enthält, die andere Knoten beauftragt (föderierte Verarbeitung). In der Analytikzone wird ein interaktives Werkzeug zur Verfügung gestellt, dass sich der eigenen Datenzone zur Datenbereitstellung bedient.
- Ein Datendienst, der als trusted research environment (TRE) dient, hat stattdessen eine offenere Programmierumgebung (zum Beispiel unter Bereitstellung von R oder Python Bibliotheken). Er hat keine eigene Datenzone, da er Daten nur anlassbezogen und temporär über die Abfragezone bezieht.

Da alle vier Arten von Knoten eine automatisierte Verarbeitung von Gesundheitsdaten auf externe Anforderung vorsehen, sind sie grundsätzlich als sichere Verarbeitungsknoten aufzusetzen, die grundlegende Sicherheitsanforderungen erfüllen müssen. Knoten mit interaktivem Nutzerzugang (Datendienste) müssen dafür weitere Sicherheitsanforderungen erfüllen.

Für die Suche und Machbarkeit dienen zum einen der Metadatenkatalog und eventuell spezielle Datendienste zur anonymen Charakterisierung von Datenbeständen. Alle Knoten mit RAZ benutzen eine zentrale Nutzeridentifikation. Die Abfragezonen kommunizieren untereinander und mit der Vertrauensstelle für die Datenverknüpfung.

Der Plattformträger überwacht die Softwarebibliothek mit den von den SPEs angebotenen und bereitgestellten Softwarekomponenten, die in den Knoten eingesetzt werden dürfen. Ein Antragsportal verwaltet die Anträge und die daraufhin erteilten Freigaben als Zugriffsrechte auf Knoten, Komponenten

und Daten. Außerdem sind alle Schnittstellen der Knoten mit entsprechenden Monitoring-Möglichkeiten ausgestattet, die durch den Plattformträger überwacht werden.

Um den Aufwand für die exekutiven Funktionen der Governance so gering wie möglich zu halten, ist die Entwicklung von Musterfällen (Use Cases) entscheidend. Wenn sich eine Anfrage bzw. ein Antrag in ein bestimmtes Muster einordnen lässt, kann ein großer Teil der Freigabeschritte automatisiert erfolgen.

Beispiel Datendienst: Ein Datendienst "patients-like-mine" (UC2) wird als separater Knoten betrieben. Alle Softwaremodule sowie Datenwege sind vorab (in einer Datenschutzfolgeabschätzung und ggf. Risikoanalyse) geprüft und zugelassen. Es ist eine Nutzerkategorie definiert, die diesen Dienst grundsätzlich nutzen darf. Eventuell kann der Datendienst auch weitere Benutzerklassen mit technischen Einschränkungen bedienen. Diese Nutzerklasse ist dann ein separater Musterfall, wird aber vom gleichen Datendienst (mit unterschiedlichen Rechteeinstellungen) bedient.

Beispiel Forschung mit Genomdaten: Ein Softwaremodul "Variantenstatistik" könnte für die GRZ vordefiniert werden. Es wird regelmäßig ausgeführt und erzeugt anonyme Daten, die dann von einem großen Teilnehmerkreis abgefragt werden können.

Beispiel Datenprojekt: Ein Forschungsprojekt möchte eine SPE mit Zugang zu klinischen Daten aus der Onkologie zusammen mit Krebsregisterdaten (UC4) haben. Hierfür kann ein Muster angelegt werden, das eventuell verschiedene Datenumfänge bereits auswählbar spezifiziert. Ein SPE-Knoten mit R-Tools kann für dieses Muster, basierend auf einem Software-Template, automatisiert angelegt werden. Erforderlich ist nur eine relativ stereotypisierte Entscheidung, ob die Antragsteller eine ausreichende Begründung für den Datenumfang angegeben haben.

4.4 Umsetzungsplan für das Modellvorhaben

Für die Umsetzung des Modellvorhabens müssen minimal folgende Funktionalitäten realisiert werden:

1. Automatisierte Datenbereitstellung aus den KDK mithilfe einer Abfragesprache, dazu ein zentraler Broker mit SPE-Schnittstelle zur Verteilung von Abfragen und Datenverknüpfung. Eine manuelle Bereitstellung, auch nur für eine Übergangszeit, erzeugt einen hohen Aufwand und konterkariert die verteilte Datenhaltung. Zusätzlich unterstützt die Funktionalität vermutlich die Datenbereitstellung zum Zweck der Evaluation des Modellvorhabens.
2. Realisierung essentieller Datendienste:
   a. Patients-like-mine (UC2) für die lernende Versorgung; erforderlich zum Nachweis des Versorgungsnutzens der Infrastruktur
   b. Qualitäts-Monitor (UC1) als Übersicht über die vorhandenen Daten und zur Machbarkeit

   Beide Datendienste nutzen die automatisierte, föderierte Datenbereitstellung.
3. Bereitstellung eines Dienstes zur Nutzeridentifizierung und Nutzerauthentifizierung insbesondere für die Versorgung. Damit können die Zugänge zu den Datendiensten freigeschaltet werden. Dieser Dienst sollte sich an externen Entwicklungen orientieren, muss aber gegebenenfalls temporär für das Modellvorhaben beim Plattformträger etabliert werden (Nutzung eines Standardtools zur Unterstützung von OAuth2 oder Äquivalent).
4. Bereitstellung eines automatisierten Dienstes zur Datenverknüpfung verschiedener Quellen durch die Vertrauensstelle, der im Rahmen der Datenbereitstellung aufgerufen werden kann. Bereitstellung einer programmierbaren SPE mit Anschluss an die GRZ und einer Bereitstellung passender Tools. Hier ist insbesondere die Umsetzung eines konkreten Use Case UC3 sicherzustellen
5. Umsetzung eines Antragsverfahrens (Etablierung eines Antragsportals, Definition der Prüfungsprozesse, Aufsetzen von Prozessen für Austausch und Rückmeldung an die Antragsteller). Erforderlich, um routinemäßig Nutzungsanträge bearbeiten zu können. Die Umsetzung sollte mit parallelen Entwicklungen (FDZ, EHDS) synchronisiert werden und soweit als möglich auf gemeinsame technische Komponenten aufsetzen.

Insbesondere die föderierte Datennutzung erlaubt es nicht, die Technik des FDZ einfach auf die Erfordernisse des Modellvorhabens zu übertragen. In dem 'Konzept zur Verknüpfung und Verarbeitung von pseudonymisierten Daten des Modellvorhabens und des Forschungsdatenzentrums', welches der Plattformträger laut gesetzlichem Auftrag im Einvernehmen mit dem Forschungsdatenzentrum und den beteiligten Vertrauensstellen dem Bundesministerium für Gesundheit bis zum 31. Mai 2026 vorlegen soll, werden diese Aspekte aufgegriffen und adressiert werden. Da dieses Konzept erst Anfang 2026 erstellt werden wird, können aktuell keine weitere Details zu diesen Verknüpfungsmodalitäten dargestellt werden.

Eine Umsetzung der Schritte 1 bis 3 sollte im Jahr 2026, die folgenden Schritte bis spätestens Ende 2027 erfolgen. Dabei ist ein Start mit eingeschränkter Funktionalität vorzusehen, der in einer weiteren Entwicklungsphase aufgrund der gemachten Erfahrungen komplettiert wird. Die genannten Funktionen sollten in einzelne Module mit standardisierten Schnittstellen zerlegt werden, sodass Entwicklung und Betrieb auf verschiedene Akteure verteilt werden können.

Aktuell wird davon ausgegangen, dass eine Realisierung der oben aufgeführten Punkte auch durch eine kombinierte Beauftragung akademischer sowie industrieller Partner umgesetzt werden könnte. Hierbei wäre auf effiziente Kommunikation zwischen den einzelnen Partnern und engen fachlichen Austausch zu achten.

5. Rahmenbedingungen und Governance

Wesentliches Ziel des Modellvorhabens ist die Erzeugung genomischer Daten von Patienten mit onkologischen oder seltenen Erkrankungen sowie die Nutzung dieser Daten für die medizinische Versorgung und Forschung. Mit dieser Stoßrichtung greift das Modellvorhaben tief in die Persönlichkeitsrechte der Betroffenen ein und gewinnt zudem eine maßgebliche forschungspolitische Tragweite. Daher muss das BfArM als Plattformträger des Modellvorhabens zeitnah dessen Rahmenbedingungen in rechtssicherer und ethisch vertretbarer Form festlegen. Dies schließt insbesondere die Schaffung einer angemessenen Governance für die im Modellvorhaben erzeugten Daten ein, verbunden mit der Klärung, wer auf welche dieser Daten unter welchen Bedingungen zugreifen und für welche Zwecke verarbeiten darf.

5.1 Datenschutz-Folgenabschätzung

Vor Inbetriebnahme der für das Modellvorhaben vorgesehenen SPE ist laut Datenschutzgesetzgebung (EU-DSGVO, Landes- und Bundesdatenschutzgesetze) von der verantwortlichen Stelle eine Datenschutz-Folgenabschätzung (DSFA) durchzuführen, um mögliche Gefährdungen der Rechte und Interessen der Datensubjekte (d.h. der Patienten und ggf. ihrer Familien) zu vermeiden bzw. zu minimieren. Die DSFA gemäß Art. 35 (7) DSGVO muss zumindest eine systematische Beschreibung der geplanten Verarbeitungsvorgänge und ihrer Zwecke sowie eine Bewertung von Notwendigkeit und Verhältnismäßigkeit in Bezug auf diesen Zweck enthalten. Des Weiteren sind eine Bewertung der Risiken für die Rechte und Freiheiten der betroffenen Personen und die zur Risikobewältigung geplanten technischer und organisatorischer Maßnahmen (TOM) - Abhilfemaßnahmen - zu dokumentieren, welche Garantien, Sicherheitsvorkehrungen und Verfahren zur Sicherstellung des Datenschutzes einschließen. Dieser Vorgang zur Risikoeinschätzung wird einmal ohne Berücksichtigung risikobezogener TOMs und einmal mit einer entsprechenden Berücksichtigung vorgenommen, sodass eine Wirksamkeitsprüfung der Maßnahmen möglich wird. Die DSFA ist im Detail durch den Plattformträger zu formulieren, sobald die genauen TOMs zur Implementierung der vorgesehenen SPEs feststehen. Dabei sind die nationalen Strategien zum Datenschutz gemäß SGB V § 64e sowie international die kommenden finalen TEHDAS2-Guidelines (speziell M7.2 "Guideline on data minimisation, pseudonymisation, anonymisation and synthetic data") zu berücksichtigen. Die TOMs sollen insbesondere das Vorgehen zur Umsetzung der Betroffenenrechte (informierte Einwilligung, Widerruf), Pseudonymisierung und Record Linkage, Datenminimierung (z.B.: Generalisierung, Suppression), Datensicherheit (z.B. Datensatzverschlüsselung) sowie Nutzungskontrolle enthalten.

5.1.1 Informierte Einwilligung der Betroffenen

Organisatorisch verantwortlich für die Einholung aller informierten Einwilligungen der Betroffenen im MV sind die Leistungserbringer, denen die qualitätsgesicherte Diagnostik und Therapiefindung mittels Genomsequenzierung im Rahmen des MV obliegt und die allein in direktem Kontakt mit den Teilnehmenden stehen. Dem Plattformträger kommt die Schlüsselrolle bei der Governance des Einholungsprozesses zu. Durch die technische Integration der Einwilligungen mit den klinischen Daten beim Klinischen Datenknoten wird gewährleistet, dass jede nachfolgende Nutzung der Daten unter Einhaltung der Einwilligungerfolgt. Die Überprüfung der Einwilligung sollte automatisiert in den Datenbereitstellungsprozess integriert werden, sodass die versehentliche Nutzung verhindert wird.

5.1.2 Pseudonymisierung

Die klinischen Daten und die Genomdaten sind unabhängig pseudonymisiert. Die Verknüpfung ist nur durch die Vertrauensstelle möglich. Um die Bezüge zwischen den Datensätzen nicht offenzulegen, wird für jede Datennutzung in der SPE eine erneute Re-Pseudonymisierung der bereitgestellten Daten durch die Vertrauensstelle vorgenommen. Auch dieser Prozess sollte automatisiert ablaufen, wenn Daten für die SPE bereitgestellt werden.

5.1.3 Datenminimierung, Datensicherheit und Nutzungskontrolle

Die föderierte Bearbeitung dient der Datenminimierung, da nur die unbedingt erforderlichen Daten anlassbezogen zusammengeführt werden. In die automatisierte Datenbereitstellung kann nur von wenigen autorisierten Personen eingegriffen werden. Durch die Bereitstellung einer spezialisierten Umgebung für jedes beantragte Projekt werden die den Nutzern zur Verfügung stehenden Verarbeitungsmittel auf das Notwendige beschränkt. Eine Protokollierung der Nutzeraktionen bei der Verarbeitung in der SPE ermöglicht dem Plattformträger eine nachträgliche Kontrolle auf angemessene Verarbeitung.

5.1.4 Technische und organisatorische Maßnahmen

Das von der EU geförderte Projekt TEHDAS II hat einen Entwurf veröffentlicht, in dem Anforderungen sowie angemessene technische und organisatorische Maßnahmen für den Betrieb sicherer Verarbeitungsumgebungen ausführlich niedergelegt sind. Dieser Entwurf soll Eingang finden in die rechtlichen Umsetzungsakte des EHDS. Die dort festgehaltenen Anforderungen sind ein geeigneter Ausgangspunkt zur Umsetzung durch den Plattformträger. Von diesen Festlegungen sollte nur in ausdrücklich begründeten Fällen abgewichen werden, um zu verhindern, dass im Rahmen des Modellvorhabens eine isolierte Infrastruktur aufgebaut wird, die dann regulatorisch nicht interoperabel mit anderen Verarbeitungsumgebungen (national/europäisch) ist. Wenn der Plattformträger Regeln festlegt, die mit anderen Verarbeitungsumgebungen nicht kompatibel sind, entstehen nahezu unüberwindliche Hürden für Datennutzungen, die Daten aus mehreren Quellen benötigen, da sie im Zweifelsfalle keine SPE finden, in der sie zusammengeführt werden dürfen (obwohl sie technisch könnten).

5.2 Governance-Modell für SPE in genomDE

5.2.1. Verantwortlichkeiten

Zu der Frage, welche Daten des MV in welcher Weise und für welche Zwecke dem Plattformträger zur Verfügung zu stellen sind, gibt es eine Reihe detaillierter Festlegungen in § 64e SGB V und der zugehörigen GenDV. Soweit Datenhaltung und Datenverarbeitung aufgrund dieser Vorgaben und im Rahmen dieser Infrastruktur erfolgen, liegt die rechtliche und organisatorische Verantwortung hierfür primär beim Plattformträger. Die Vertrauensstelle ist für die von ihr gesteuerten Verarbeitungsvorgänge ebenfalls Verantwortlicher im Sinne des Datenschutzrechts. Für den einleitenden Übermittlungsvorgang in die Plattform liegt die Verantwortlichkeit bei den Leistungserbringern, für die abschließende Nutzung

bei den Nutzungsberechtigten – ggf. in Form einer gemeinsamen Verantwortlichkeit mit dem Plattformträger für die Verarbeitung innerhalb der SPE. Demgegenüber fungieren die Genomrechenzentren, klinischen Datenknoten und Datendienste als bloße Auftragsverarbeiter.

5.2.2. Nutzungsberechtigungen

Die Nutzungsberechtigung für Daten, die im Rahmen des MV erhoben und gespeichert werden, ergeben sich ebenfalls weitgehend aus den Vorgaben des § 64e SGB V und der GenDV. Sie sind nicht mehr akteurs-, sondern zweckbezogen definiert. Nutzungsberechtigt ist jede natürliche oder juristische Person im Anwendungsbereich der DSGVO, unabhängig davon, ob sie selbst am MV teilnimmt oder nicht. Zulässige Nutzungszwecke sind die Gesundheitsversorgung, Qualitätssicherung, Evaluation und Forschung. Für eine Forschungsnutzung oder eine Fallidentifizierung zu Versorgungszwecken ist eine eigenständige Einwilligung der teilnehmenden Versicherten erforderlich – ansonsten besteht auf der Basis der generellen Einwilligung in die Genomsequenzierung und in die Teilnahme am MV eine gesetzliche Verarbeitungsbefugnis aller Plattformakteure. Die Forschungseinwilligung wird nicht spezifisch für das MV eingeholt; stattdessen wird dafür auf die in der Praxis bereits etablierten Forschungseinwilligungen insb. die der MII zurückgegriffen.

Prozedural setzt jede Nutzung der in der Plattform gespeicherten Daten einen Antrag bei dem, und eine Freigabe durch den, Plattformträger heraus. Dieser setzt zur Bewältigung des Verfahrens im Sinne eines Use and Access Committee den dafür gesetzlich vorgesehenen wissenschaftlichen Beirat ein.

Die Governance-Regeln der Plattform gelten nur für die Datenverarbeitung *unter Nutzung der Plattforminfrastruktur*. Die Leistungserbringer behalten die Möglichkeit, die von ihnen eingespeisten Daten parallel in anderen Systemen zu speichern und dort eigenständig für Versorgungs-, Qualitätssicherungs- oder Forschungszwecke zu verwenden – etwa auf der Grundlage der Forschungseinwilligung der MII.

5.2.3. Bereitstellungsverfahren

Die Bereitstellung der Daten des MV erfolgt über die in § 64e SGB V und GenDV dafür vorgesehenen Einrichtungen, d.h. Treuhandstelle, Genomrechenzentren, Klinische Datenknoten und Datendienste. Detaillierte Vorgaben zur entsprechenden Rollenverteilung gibt es hierbei für die Wiederherstellung des Fallbezuges. Für die Bereitstellung pseudonymisierter Einzeldatensätze wird die Datennutzung in SPEs zwingend vorgeschrieben, deren Ausgestaltung wiederum Gegenstand des von genomDE vorgelegten Konzeptes ist. Im Übrigen ist die technische Ausgestaltung der Datenbereitstellung rechtlich nicht im Einzelnen vorgezeichnet und muss von den beteiligten Akteuren festgelegt werden. Nicht klar verortet ist bislang die Prüfung bestimmter Nutzungsvoraussetzungen v.a. im Forschungsbereich, nämlich Vorliegen und Umfang einer diesbezüglichen Einwilligung sowie das Vorliegen eines Ethikvotums.

5.2.4. Handlungsempfehlungen

Jede Form einer Institutionalisierung der Governance der Dateninfrastruktur des MV, einschließlich deren SPEs, wird sich – jenseits der Regelungen in § 64e SGB V – an nationalen (GDNG) und internationalen (EHDS) gesetzlichen Vorgaben und Rahmenbedingungen orientieren müssen. Im Zentrum des Interesses sollte dabei jedoch Funktionalität, Transparenz und Nachhaltigkeit der Governance stehen. Insbesondere ist zu berücksichtigen, dass die Governance nach Abschluss des MV einer Ausweitung der genommedizinischen Versorgung auf weitere Krankheitsentitäten und Leistungserbringer gerecht werden muss. Deswegen ist zu empfehlen, mögliche Protagonisten einer solchen Erweiterungen frühzeitig in die Planung der künftigen Governance einzubinden.

Generell bieten die derzeitigen gesetzlichen Regelungen für alle avisierten Nutzungsszenarien eine hinreichende rechtliche und organisatorische Basis. Kleinere Korrekturen sind jedoch zu empfehlen:

- Die verbliebenen Einwilligungserfordernisse in § 64e Abs. 6 S. 2 SGB V sollten überdacht werden. Soweit die Nutzungsberechtigten infolge der vorgeschriebenen Verwendung von SPEs gar keine

personenbezogenen Daten mehr erhalten, ist – jenseits der generellen Einwilligung in die Teilnahme am MV – ein besonderes Einwilligungserfordernis für Forschungszwecke oder eine Fallidentifizierung normativ nicht geboten.
- GenDV Anlage, Abschnitt III Ziff. 2 sollte dahingehend ergänzt werden, dass Informationen zu Vorliegen und Umfang einer Forschungseinwilligung an die Genomrechenzentren und klinischen Datenknoten übermittelt werden. Zudem sollten die Prüfungszuständigkeiten für das Vorliegen der Einwilligung und eines Ethikvotums festgelegt werden. Soll diese Prüfungszuständigkeit dem Plattformträger zugewiesen werden, müssten die dafür erforderlichen Informationen von Genomrechenzentren und klinischen Datenknoten verpflichtend an den Plattformträger übertragen werden. Die Notwendigkeit der dann erforderlichen doppelten Haltung patientenspezifischer Daten sollte jedoch sorgsam gegen das Gebot der Datensparsamkeit abgewogen werden.
- Die gesetzliche Übergangsklausel in § 64e Abs. 13 S. 2 SGB V sollte so überarbeitet werden, dass sie keine automatische und unmittelbare Beendigung des MV mit Vorlage des Sachverständigenberichts bewirkt, auch nicht bei einem negativen Votum, sondern zumindest die Fortführung der Plattformfunktionalitäten bis zu einer Entscheidung des Gesetzgebers sicherstellt.
- Die Forschungseinwilligung der MII sollte für die Zukunft so angepasst werden, dass sie eine mögliche Übertragung von Entscheidungsbefugnissen zur Forschungsnutzung der Daten des Patienten bzw. der Patientin vom Versorger an geeignete Dritte – wie im vorliegenden Fall den Plattformträger des MV – klar abdeckt.
- Zudem sollte erwogen werden, den MII-Mustertext um die Klarstellung zu ergänzen, dass eine Löschung nach Widerruf der Forschungseinwilligung auch dann unterbleibt, wenn es weitere zulässige Nutzungszwecke gibt.

Ungeklärt ist nach wie vor das weitere Vorgehen nach Beendigung des MV. Die aktuellen gesetzlichen Regelungen sehen eine vorübergehende Fortsetzungsmöglichkeit durch Selektivverträge vor. Angesichts der praktischen und rechtlichen Tragweite bedarf es hier aber vorausgreifender Überlegungen und einer möglichst zeitnahen und breiten Abstimmung, wie die Plattform und ihre Funktionen in den existierenden Rechtsrahmen eingefügt werden können. Es wäre also zu klären, wem nach Beendigung des MV welche der derzeitigen Aufgaben, Rechte und Pflichte der Plattformteile in welcher Form übertragen werden sollen, und wie dies in die größeren Strukturen des EHDS und des GDNG einzubinden ist.

5.3 Governance-Prozessmodell

Die Governance muss an drei Stellen reaktiv organisiert sein:

a) Bezogen auf die Anforderungen und Wünsche der Teilnehmer (Rechtspersonen: KDK, GRZ, Betreiber von Datendiensten)
b) Bezogen auf die Anforderungen und Wünsche der Nutzer (Natürliche Personen: Heilberufe, Forschende, Firmenvertreter, Bürger)
c) Bezogen auf die Anforderungen und Wünsche der Patientinnen und Patienten

Dafür müssen Prozesse institutionalisiert und die notwendigen technischen Werkzeuge bereitgestellt werden. Beides muss kontinuierlich weiterentwickelt werden.

5.3.1 Governance der Teilnehmer

Teilnehmer sind Rechtspersonen, die Leistungen im Rahmen der Infrastruktur erbringen. In der Regel schließen diese Verträge mit dem Plattformträger zur Festlegung der Rechte und Pflichten.

Grundsätzliches kann in einer Teilnahmeordnung bzw. in Musterverträgen geregelt sein. Hier können auch Anforderungen bezüglich der erforderlichen technischen und organisatorischen Maßnahmen (TOM) niedergelegt werden.

Die folgenden Aktivitäten müssen abgedeckt werden:

- Onboarding/Offboarding
- Zulassung von Funktionen
- Monitoring

Das Onboarding umfasst den Abschluss der Vereinbarungen.

Die Zulassung von Funktionen betrifft sowohl den Betrieb von vernetzten Knoten als auch den Betrieb spezifischer Softwarefunktionen innerhalb eines Knotens. Ein Teilnehmer kann sich auch der Auftragsdatenverarbeitung mit Dritten bedienen. Funktionen in puncto Secure Processing betreffen komplette Datendienste sowie Softwaremodule, die im Software-Repository zur Nutzung aufgenommen werden sollen. Die Zulassung beinhaltet eine Prüfung der verwendeten Software und der technischen Konfiguration.

Das Monitoring betrifft kontinuierliche (IT-Monitoring) und regelmäßige (Begehungen, Zertifizierung, Akkreditierung) Überprüfungen auf Vertragspflichten.

5.3.2 Governance der Nutzeranfragen

Nutzer sind natürliche und juristische Personen, die ein legitimes Anliegen zur Datennutzung haben.

Im Rahmen eines Antragsprozesses (sei es einmalig für den wiederkehrenden Zugang zu einem existierenden Datendienst, sei es einmalig für eine spezifische Projektkonfiguration) müssen die folgenden Aktivitäten abgedeckt werden:

- Auffindbarkeit
- Machbarkeit
- Genehmigung
- Bereitstellung
- Nachbereitung

Grundsätzlich muss ein Mechanismus zur Identifizierung und Authentifizierung von Nutzern bereitgestellt werden.

Auffindbarkeit betrifft die Organisation eines Metadatenkataloges (inhaltlich gespeist von den Teilnehmern und strukturiert durch die wissenschaftliche Nutzergemeinschaft).

Machbarkeit betrifft die Bereitstellung (durch Teilnehmer bzw. wissenschaftliche Gemeinschaft) von Machbarkeitswerkzeugen und Portalen. Die Werkzeuge müssen geprüft werden (s. "Zulassung von Funktionen" weiter oben).

Genehmigung betrifft die Entgegennahme eines Antrags in einem Antragsportal sowie dessen Prüfung und Genehmigung. Der Antrag umfasst insbesondere eine Zweckbestimmung (als Freitext mit zusätzlicher Kategorisierung in gewisse Musterfälle) sowie eine Datenkonfiguration (als technische Spezifikation des gewünschten Datenumfangs via Query sowie als strukturierte Spezifikation der Werkzeuge und Ressourcen via "Häkchenliste"). Erforderliche Prüfungen können auch an fachlich qualifizierte Teilnehmer (aka verteilte HDABs) delegiert werden.

Die Bereitstellung umfasst die technische, sicher signierte Weiterleitung der genehmigten Spezifikation an die entsprechenden Knotenbetreiber (Autorisierung für einen Datendienst bzw. Einrichtung eines Datenprojekts in einem TRE) sowie ggf. an die Vertrauensstelle (Autorisierung). Nutzerseitig beigestellte Tools und Daten müssen geprüft und ebenfalls übermittelt werden.

Die Nachbereitung umfasst die Prüfung ggf. ausgeleiteter Daten eines Projekts sowie die Nachverfolgung von Veröffentlichungs- und Löschpflichten.

5.3.3 Governance der Patientenanliegen

Aktuellen rechtlichen und ethischen Vorgaben folgend sind bei der Ausgestaltung der Governance der IT-Infrastruktur des MV die Interessen der Betroffenen zu berücksichtigen. Insbesondere sind diese über Art, Umfang und Ergebnisse der Nutzung ihrer Daten in geeigneter Form zu informieren. Da im Absatz 11b des §64e SGB V vorgesehen ist, dass die entsprechenden Informationen ohnehin von allen Datennutzern an den Plattformträger übermittelt werden müssen, fällt die Aufgabe einer Zugänglichmachung für die Patienten damit primär dem Plattformträger zu. Idealerweise sollte hierfür ein Informationsportal eingerichtet werden, dass sich an vergleichbaren Entwicklungen im Rahmen der Medizininformatik-Initiative oder vergleichbaren Dateninitiativen orientiert.

Daneben ist die Einhaltung aller einschlägigen datenschutzrechtlichen Vorgaben zur Kontrolle der Betroffenenrechte zu gewährleisten. Dies erfordert insbesondere ein von Plattformträger und Treuhandstelle zu betreibendes und verantwortendes Einwilligungsmanagement, in dem alle allfälligen Änderungen der eingeräumten Nutzungsrechte verlässlich und zeitnah abgebildet und umgesetzt werden. Soweit möglich sollte dabei auf bereits existierende Lösungen zurückgegriffen werden, wie sie z.B. im Rahmen der Medizininformatik-Initiative entwickelt wurden.

Zur Weiterentwicklung und Kontrolle der Einhaltung der Patientenrechte sieht Absatz 9a des §64e SGB V ausdrücklich die Repräsentanz der „maßgeblichen Bundesorganisationen für die Wahrnehmung der Interessen von Patientinnen und Patienten und der Selbsthilfe chronisch kranker Menschen" im Beirat des MV vor.

5.4 Vorschlag Data Governance

Um den Plattformträger von der fachlichen und inhaltlichen Beurteilung von Datenzugangsanträgen zu entlasten, wird die Einrichtung eines Gremiums in Anlehnung an gängige Use and Access Committees (UAC) vorgeschlagen[5]. Die Aufgaben des UAC können durch eine entsprechend §64e SGB V Absatz 9a neu eingerichtete Arbeitsgruppe oder den in Absatz 9b vorgesehenen Wissenschaftlichen Ausschuss übernommen werden.[6]

Das UAC sollte zwischen 5 und 8 Mitglieder umfassen, die wie üblich gemäß eines vom BMG festgelegten Schlüssels durch den Plattformträger berufen werden. Bei der Festlegung des Schlüssels sollte neben der Repräsentanz der notwendigen fachlichen Expertise (Medizin, IT, Ethik, Datenschutz) auch die Einbindung von Patienteninteressen berücksichtigt werden.

Kern des Verfahrens ist eine Leitung des UAC in Form einer Doppelspitze, die von einer bzw. einem ehrenamtlichen Vorsitzenden und einer hauptamtlichen Geschäftsführung der Geschäftsstelle des UAC gebildet wird. In Abhängigkeit von der tatsächlich anfallenden Arbeitslast kann es erforderlich werden, die bzw. der Vorsitzenden auf Kosten des Modellvorhabens teilweise oder ganz von seinen anderen beruflichen Aufgaben freizustellen.

[5] Die Formulierung des vorliegenden Vorschlags orientiert sich an der Geschäftsordnung des UAC des Universitätsklinikums Schleswig-Holstein UKSH, die laut Vorstandsbeschluss vom Juni 2025 am 1. Januar 2026 in Kraft tritt. Das geschilderte Verfahren hat sich im UKSH Campus Kiel langjährig bewährt und wurde im Sommer 2025 per Beschluss des Vorstandes mit Wirkung zum 1.1.2026 auch für den Campus Lübeck beschlossen. Der Plattformträger sollte in Anlehnung an die zugehörige Geschäftsordnung eine entsprechende Regelung auch für die Data Governance des Modellvorhabens festlegen.

[6] Es wird jedoch davon abgeraten, die Aufgaben eines UAC an einen wissenschaftlich hochkarätig besetzten Beirat oder Ausschuss zu übertragen, da erfahrungsgemäß die Bereitschaft zur routinemäßigen Übernahme solcher Aufgaben schnell erlahmen kann. Diese Gefahr besteht umso mehr, je umfangreicher und durch weitere Aufgaben belastet ein solches Gremium ist.

Die Doppelspitze entscheidet einvernehmlich, ob ein Antrag an das UAC bereits festgelegten Kriterien genügt und daher unmittelbar genehmigt werden kann (einfaches Verfahren), oder ob sich das UAC in einer seiner regelmäßigen Sitzungen mit dem Antrag befassen muss (erweitertes Verfahren). Über Entscheidungen im einfachen Verfahren wird das UAC von der Doppelspitze zeitnah und regelmäßig informiert. Den Mitgliedern steht hierbei ein Vetorecht zu, dessen Wahrnehmung eine Antragsbefassung im erweiterten Verfahren bewirkt. Der Ausschuss gibt Empfehlungen an den Plattformträger ab. Der Plattformträger schließt sich üblicherweise diesen Empfehlungen an.

5.5 Vorschlag Teilnehmer Governance

Um den Plattformträger von detaillierten technischen Beurteilungen zu entlasten, ist es sinnvoll, die Anforderungen an die verschiedenen Teilnehmergruppen in Form von Richtlinien festzulegen, mit denen Teilnehmer sich dann in Eigenleistung unabhängig zertifizieren oder akkreditieren lassen können. Dies gilt insbesondere für die Anforderungen an Datendienste und allgemeine SPE.

Solche Richtlinien existieren bereits an einigen Stellen (zum Beispiel am BSI), andere Richtlinien sollten sich sinnvollerweise an aktuellen Entwicklungen (zum Beispiel des EHDS) orientieren. Die Erarbeitung der Richtlinien sollte in einem Community-Prozess erfolgen und nicht allein auf der Hoheit des Plattformträgers basieren.

Mehr noch als die technische Interoperabilität bedarf die regulatorische Interoperabilität einer sorgfältigen Festlegung und kontinuierlichen Anpassung der Zulassungskriterien für Teilnehmer. Es ist sinnvoll, dafür eine nationale Institution zur Harmonisierung dauerhaft zu etablieren.

5.5.1 Äquivalenz von SPEs

Sichere Verarbeitungsknoten unter der Ägide des Modellvorhabens müssen mit anderen SPE zusammenarbeiten können. Ein Forschungsprojekt, das auf der einen Seite eine Analyse vollständiger genomischer Daten erfordert und auf der anderen Seite KI-basierte Auswertungen von Bildern verwendet, muss Daten aus zwei unterschiedlichen SPE integrieren können. Die technischen Schnittstellen sind dabei tatsächlich weniger wichtig, weil sie im Rahmen eines umfangreichen Projektes auch angepasst und weiterentwickelt werden können. Viel wichtiger ist die regulatorische Interoperabilität: wenn Daten nicht in ein anderes SPE ausgeleitet werden können, weil diesem nicht vertraut wird oder es rechtlich nicht möglich ist (unterschiedliche Zulassungskriterien), ist in der Regel keinerlei kurzfristige Abhilfe möglich. Dies gilt sowohl für nationale Knoten, die auf unterschiedlichen Rechtsgrundlagen basieren, als auch für die zukünftige internationale Zusammenarbeit.

5.5.2 Verknüpfbarkeit von Datenquellen

Voraussetzung für eine übergreifende Datennutzung ist ein einheitlicher Prozess, Daten über Datenquellen hinweg zu verknüpfen. Neben der eigentlichen Record Linkage ist dafür auch eine Abstimmung der entsprechenden Rechtsgrundlagen erforderlich.

5.5.3 Institutionalisierung

Die Erarbeitung gemeinsamer Richtlinien erfordert eine große Nähe zu den Datengebern, den Datendienstleistern und den Nutzern aus Versorgung, Forschung und Industrie. Auch eine Interessenvertretung in Richtung der Gesetzgebung ist wünschenswert. Wir schlagen daher vor, eine Institution zu schaffen, die von der oben genannten Community getragen wird. Bisherige klinisch-medizinisch orientierte Netzwerke sind zu begrenzt und nicht nachhaltig verfasst.

Konkrete Aufgaben der Institution sind: Erarbeitung und Abstimmung von Richtlinien und Kriterien für Governance-Prozesse (Zertifizierung, Zulassungskriterien, Monitoring-Ansätze), Geschäftsmodelle, internationale Einbindung.

5.6 Entgelte und Vergütung für die Nutzung der SPE-Dateninfrastruktur

Für das Modellvorhaben können drei Finanzierungsnotwendigkeiten unterschieden werden:

1. Der Aufbau und der Betrieb der Dateninfrastrukturaufbau mit den einzelnen Datenknoten wird durch den Plattformträger BfArM mit Mitteln aus dem Bundeshaushalt finanziert.
2. Die Datenerhebung, Verarbeitung, Qualitätsprüfung, Speicherung in der Dateninfrastruktur werden durch Mittel der gesetzlichen und teilnehmenden privaten Krankenversicherungen finanziert, mit denen die Leistungserbringer vergütet werden. Geregelt ist dies im Vertrag[7] zwischen dem GKV-Spitzenverband und den Leistungserbringern. Diese Vergütung aus den gesetzlichen und teilnehmenden privaten Krankenkassen endet mit dem Vorgang der qualitätsgesicherten Übermittlung der Daten an die KDK und GRZ.
3. Für die Nutzung der Daten in SPEs, insbesondere in der medizinischen Forschung, steht noch ein Vergütungsmodell aus. Der grundsätzliche Rahmen für ein solches Vergütungsmodell soll hier beleuchtet werden.

Für die Entwicklung eines Vergütungsmodells für die SPE-Nutzung ist es sinnvoll, die verschiedenen Kostenarten zu unterscheiden, die bei einer Datennutzung relevant werden. Die Kosten der Datennutzung können einen Anhaltspunkt für mögliche Positionen in einem Entgeltkatalog für die Datennutzung geben. Da es auch einen Markt für Forschungsdaten gibt, sind Kosten als Bezugsgröße für Entgelte immer auch mit den Marktpreisen für ähnliche Leistungen abzuwägen.

Da die Daten des Modellvorhabens finanziert und qualitätsgesichert in den Datenknoten vorliegen und auch deren Betrieb durch Bundesmittel gesichert ist, liegt der Fokus für die Herleitung von Entgelten auf allen Leistungen, die für das Antragsverfahren, der Prüfung und Zurverfügungstellung von Daten in SPEs benötigt werden und die dafür notwendige Infrastruktur. Dabei werden projektspezifische (u.U. flexible) und regelmäßige Kosten entstehen.

Die regelmäßigen Kosten entstehen durch die notwendige Vorhaltung von Infrastrukturen beim Plattformträger und bei den SPE-Betreibern. Regelmäßige, "fixe" Kosten entstehen zu großen Teilen durch Personalaufwand, der beim Plattformträger zur Bearbeitung der nicht automatisierten Teile des Antragsverfahren, der Überwachung der Nutzung sowie der Überwachung der Ergebnisausleitung anfällt. Bei den SPE-Betreibern wird Personalaufwand für die Koordination und den technischen Support der Datennutzung anfallen. Der Aufbau eines entsprechend qualifizierten Personalressource ist langfristig zu planen und umzusetzen. Personalkosten sind daher nur wenig flexibel.

Weitere regelmäßig anfallende Kosten entstehen durch die Vorhaltung der notwendigen Technik und Ausstattung. Die technische Ausstattung für die Datennutzung in SPEs wird absehbar schnellen Entwicklungszyklen unterliegen und sehr regelmäßig neu investiert werden müssen. Gleichzeitig ermöglicht der schnelle Entwicklungszyklus u.U. auch Effizienzvorteile, die sich in den projektspezifischen Kosten abbilden können. Die weitere Ausstattung z.B. von Arbeitsplätzen tritt im Vergleich mit den Personalaufwänden in den Hintergrund und kann pauschal pro Arbeitsplatz geschätzt werden.

Projektspezifische Kosten fallen für Leistungen an, die durch Projekte selbst ausgelöst werden. Dies betrifft alle Leistungen, die für ein Projekt zwischen erster Beratung durch den Plattformträger, Antragstellung, Prüfungsprozess, Aufbau der SPE-Instanzen und Verfügbarmachung von Ergebnissen anfallen. Die Leistungen können je nach Art und Komplexität der Datennutzung sehr unterschiedlich sein. Ihre Kosten sollten jeweils abgeschätzt werden. Die Kosten für die Entwicklung der Verfahren, z.B. des Antragsverfahrens beim Plattformträger, sind eher den Vorhaltekosten und damit der Finanzierung durch den Bund zuzurechnen.

[7] "Vertrag zur Durchführung eines Modellvorhabens zur umfassenden Diagnostik und Therapiefindung mittels einer Genomsequenzierung bei Seltenen und bei onkologischen Erkrankungen nach § 64e SGB V zwischen dem GKV-Spitzenverband, Berlin und den Leistungserbringern" aus dem Frühjahr 2024.

Leistungen könnten folgendermaßen bewertet werden (Liste ist nicht abschließend):

Leistung	**Kosten-zuordnung**	**Kostenarten**	**Kostenträger**	**Empfehlung Finanzierung**
Beratung im Antragsverfahren	projektspezifisch	Personalkosten (spezifisch) Overhead (pauschal) Sachkosten (pauschal)	BfArM	ohne Entgelt für akademische Forschung
Machbarkeitsprüfung	projektspezifisch	Personalkosten (spezifisch) Overhead (pauschal) Sachkosten (pauschal)	BfArM	spezifisches Entgelt, Bemessungsfaktor zu definieren
Machbarkeitsprüfung in Dash Board	Vorhaltung	Sachkosten (pauschal)	BfArM	Bund
Prüfung des Antrages	projektspezifisch	Personalkosten (spezifisch) Overhead (pauschal) Sachkosten (pauschal)	BfArM	pauschales Entgelt
Vorhaltung Technik und Infrastruktur SPE	Vorhaltung	Sachkosten (pauschal)	BfArM	Bund
automatisierter Aufbau der SPE-Instanz	projektspezifisch	Sachkosten (spezifisch)	SPE-Betreiber	ohne Entgelt (automatisiert)
möglicherweise manueller Aufbau der SPE-Instanz	projektspezifisch	Personalkosten (spezifisch) Overhead (pauschal) Sachkosten (pauschal)	SPE-Betreiber	spezifisches Entgelt, Bemessungsfaktor Zeitaufwand
CPU-Zeit	projektspezifisch	Sachkosten (spezifisch)	SPE-Betreiber	spezifisches Entgelt, Bemessungsfaktor Technikkosten/Zeit
Transferbedarf	projektspezifisch	Sachkosten (spezifisch)	SPE-Betreiber	spezifisches Entgelt, Bemessungsfaktor Datengröße
Speicherkapazität, Fallbasierte Vorhaltung, etc.	projektspezifisch	Sachkosten (spezifisch)	SPE-Betreiber	spezifisches Entgelt, Bemessungsfaktor Speicherplatz
Data Stewardship	Vorhaltung	Personalkosten (spezifisch) Overhead (pauschal) Sachkosten (pauschal)	Datenknoten und SPE-Instanzen	Bund
Ergebnisprüfung (automatisiert, plausibilisiert durch Menschen)	projektspezifisch	Personalkosten (spezifisch) Overhead (pauschal) Sachkosten (pauschal)	BfArM	pauschales Entgelt

Leistung	Kosten-zuordnung	Kostenarten	Kostenträger	Empfehlung Finanzierung
Ergebnisprüfung und händische Datenbereitstellung	projektspezifisch	Personalkosten (spezifisch) Overhead (pauschal) Sachkosten (pauschal)	BfArM	spezifisches Entgelt, Bemessungsfaktor Zeitaufwand

Der Plattformträger sollte in einem einheitlichen Entgeltkatalog die spezifischen und pauschalen Entgelte festlegen, die dann für alle Datennutzer der Daten des Modellvorhabens gelten. Diese Entgelte müssen gegenüber den Datennutzern so gesetzt werden, dass Projekte möglich werden können (keine Kosten für Beratung, wenn noch keine Finanzierungszusagen bei akademischen Projekten vorliegen) und gleichzeitig die Kosten nicht so hoch sind, dass Projekte auf andere gleichwertige Datenquellen ausweichen. Ob zwischen privat und öffentlich organisierten Datennutzern bei der Entgeltbemessung unterschieden werden soll, ist politisch zu beantworten[8].

Ein einheitlicher Entgeltkatalog sollte gleichzeitig bedeuten, dass die Leistungen für alle gleichartigen Infrastrukturelemente (z.B. alle KDK) gleich bewertet werden. Für gleichartige Leistungen werden entsprechend Pauschalen vergütet. Dies vereinfacht das intern notwendige Verrechnungssystem wesentlich und ermöglicht eine schnelle Kalkulation der projektspezifischen Entgelte. Einzelverhandlung mit jedem Datenknoten bei jedem Projekt sind damit ausgeschlossen. Gleichzeitig müssen die Kosten in einer ersten Abrechnungsperiode für alle Infrastrukturelemente kostendeckend sein. Um die Entgelte (und die Forschungsbudgets) im Rahmen zu halten, könnte in späteren Kalkulationsperioden ein Effizienzfaktor für bestimmte Infrastrukturelemente eingesetzt werden. Dieser würde durch die Bestimmung eines Anpassungspfades für pauschale Kostenvergütungen ineffiziente Infrastrukturelemente unter Druck setzen und effiziente Infrastrukturelemente belohnen. Ein Beispiel für ein entsprechendes Verfahren ist die Regulierung der Energienetze in Deutschland[9]. Die pauschalen Vergütungen für Leistungen der Infrastrukturelemente als Basis für die Entgelte sollten regelmäßig spätestens alle 3 Jahre neu verhandelt werden.

Für die Machbarkeitsprüfung übermittelt der Plattformträger die Anforderungen des geplanten Projektes an die SPE-Betreiber. Diese sind dann für die Abschätzung der spezifischen Leistungsanteile und die damit gemäß Entgeltkatalog anfallenden spezifischen Entgeltanteile zuständig und melden diese zurück an den Plattformträger. Zur Einschätzung der Kosten einer föderierten Datenverarbeitung holt jeder SPE-Betreiber einen Kostenvoranschlag gemäß Entgeltkatalog von dem jeweiligen Auftragsdatenverarbeiter ein. Der Plattformträger trägt in erster Instanz die Kosten und gibt diese gemäß Entgeltkatalog an die Forschenden weiter. Da die Kostenschätzung Teil der Machbarkeitsanalyse ist, ist diese Planung durchzuführen, bevor eine Freigabe des Projekts erfolgt. Datenknoten sind deshalb angehalten, für solche Use-Cases einen Verarbeitungsplan bereitzuhalten.

Aus der Erfahrung mit der Entgeltherleitung in anderen Projekten mit einer föderierten Infrastruktur, die durch verschiedene Institutionen betrieben wird, empfehlen wir den oben dargestellten Weg der Aushandlung von Pauschalen für Leistungen, die in absehbarer Zeit mit einem Effizienzziel hinterlegt werden.

[8] Eine Orientierung für Institutionen, für die ein gemindertes Entgelt berechnet werden könnte, findet sich in der Datentransparenz-Gebührenverordnung - DaTraGebV §11 Absatz 3 und 4 zum FDZ-Gesundheit.
[9] https://www.bundesnetzagentur.de/DE/Fachthemen/ElektrizitaetundGas/Netzentgelte/Anreizregulierung/WesentlicheElemente/Effizienzwert/start.html

6. Ausblick

SPEs werden in vielen Kontexten benötigt und aktuell konzipiert. Sie sind ein entscheidender Baustein, um die Vorteile der Digitalisierung im Gesundheitswesen wie z.B. effizientere Versorgung und Forschung sowie evidenzbasierte Gesundheitspolitik mit dem Schutz personenbezogener (und deshalb besonders sensibler) Daten zu vereinbaren.

Auch außerhalb des Gesundheitssektors sind SPEs auf dem Vormarsch. Sie entwickeln sich von spezialisierten, projektbezogenen „Schutzzonen" zu robusten, flexiblen und kommerziell nutzbaren Datenverarbeitungs-Infrastrukturen.

Die Treiber dieser Entwicklung sind

- steigende Sicherheits- und Datenschutzanforderungen,
- technologische Innovationen (KI, PETs: Privacy-Enhancing Technologies),
- Nachfrage nach vertrauenswürdiger Datenverarbeitung,
- wachsende Datenmengen in sensiblen Bereichen und die Notwendigkeit ihrer Verknüpfung.

In den kommenden Jahren werden SPEs daher zu einem allgemeinen Standard der sicheren Datenanalyse, nicht nur im Gesundheitsbereich. Interoperabilität, angemessene Nutzungskosten und Nutzerfreundlichkeit werden neben der erreichten Datensicherheit die entscheidenden Erfolgskriterien für SPEs sein. Daneben entscheiden verlässliche, schnelle und niederschwellige Prozesse der Datenbereitstellung in den SPE maßgeblich über die Akzeptanz des SPE-Ansatzes allgemein.

Das BfArM übernimmt als Plattformträger die Aufgabe, die Daten des Modellvorhabens in SPEs zugänglich zu machen. Erste Anträge auf Datennutzung und Datendienste sind bereits gestellt, so dass sich der Plattformträger zeitnah der Umsetzung des hier vorgelegten Konzeptes wird annehmen müssen.