Virtuelle Forschungsumgebung für Gesundheitsdaten: Virtual Research Environment (VRE) und Health Data Cloud (HDC)

Michael Schirner[1,2,3,4,5] *und Petra Ritter*[1,2,3,4,5]

[1]Berlin Institute of Health at Charité, Universitätsmedizin Berlin, Charitéplatz 1, Berlin 10117, Germany

[2]Department of Neurology with Experimental Neurology, Charité, Universitätsmedizin Berlin, Corporate member of Freie Universität Berlin and Humboldt Universität zu Berlin, Charitéplatz 1, Berlin 10117, Germany

[3]Bernstein Focus State Dependencies of Learning and Bernstein Center for Computational Neuroscience, Berlin, Germany

[4]Einstein Center for Neuroscience Berlin, Charitéplatz 1, Berlin 10117, Germany

[5]Einstein Center Digital Future, Wilhelmstraße 67, Berlin 10117, Germany

I. Einleitung

Die zunehmende Nutzung digitaler Daten erfordert die Entwicklung robuster und kollaborativer digitaler Plattformen, die Datensicherheit, Datenschutz und die Einhaltung rechtlicher Rahmenbedingungen gewährleisten.[1] Medizinische Forschung und insbesondere die Entwicklung von künstlicher Intelligenz für Diagnose und Therapie macht die Verarbeitung von Gesundheitsdaten zwingend erforderlich. Auf ähnliche Weise ist beispielsweise das Feld Robotik im Kontext der Entwicklung sogenannter Digital Twins betroffen: um simulierte Ebenbilder von Menschen zu entwickeln, müssen Daten verarbeitet werden die die Gesundheit der abgebildeten Menschen charakterisieren. Das Virtual Research Environment (VRE)[2] bil-

1 https://www.eneuro.org/content/10/2/ENEURO.0215-22.2023/tab-article-info.
2 vre.charite.de.

det die Grundlage der Health Data Cloud (HDC)[3] und ist eine digitale Forschungsplattform, die den Prinzipien des European Health Data Space (EHDS)[4] und der Datenschutzgrundverordnung (DSGVO)[5] zum Schutz persönlicher Daten folgt, aber Forschern gleichzeitig die nötige Flexibilität für die Implementierung eigener Verfahrensabläufe einschließlich experimenteller Verarbeitungsverfahren bietet. Die VRE/HDC unterstützt Verantwortliche und Auftragsverarbeitende Datenschutz durch Technikgestaltung und durch datenschutzfreundliche Voreinstellungen zu praktizieren und Compliance mit Datenschutzvorgaben zu demonstrieren. Die VRE/HDC ist vollständig quelloffen und basiert auf weitläufig benutzten Open-Source-Paketen, um sichere und interoperable Gesundheitsforschung zu ermöglichen. Die VRE/HDC bietet optimierte technische und organisatorische Maßnahmen (TOMs) für die geschützte Verarbeitung und kontrollierte gemeinsame Nutzung großer Datensätze, einschließlich biomedizinischer Bildgebungsdaten und digitaler Zwillingsmodelle (digital twins) des Menschen.

In diesem Artikel skizzieren wir einige Maßnahmen, wie VRE/HDC die Einhaltung der DSGVO Prinzipien umsetzen. Wir skizzieren auch den Ansatz von VRE/HDC zur Datenstandardisierung, Reproduzierbarkeit für die Umsetzung der FAIR-Datengrundsätze, einschließlich der Versionierung von Datensätzen, der Annotation von Metadaten und der Registrierung in einem durchsuchbaren Knowledge Graph, standardisierter Formate für Daten und Metadaten mit interoperabler ontologischer Darstellung des in Datensätzen enthaltenen Wissens.

Forschende arbeiten häufig in einer abgetrennten digitalen Umgebung, zum Beispiel an einem Laptop mit einem eigenen Dateisystem, was den kontrollierten Austausch sensibler Daten erschwert. Ein weiteres Problem ist, dass die verfügbaren Ressourcen eines persönlichen Computers im Vergleich zu Hochleistungsrechnern nicht nach Bedarf skaliert werden können. Dies ist häufig nötig, beispielsweise nachdem ein experimentelles Verarbeitungsverfahren fertig gestellt wurde, um es danach automatisiert auf einen großen Datensatz anzuwenden. Ein weiterer Nachteil ist, dass mit solchen privaten Datenräumen die Daten nicht leicht mit anderen gemein-

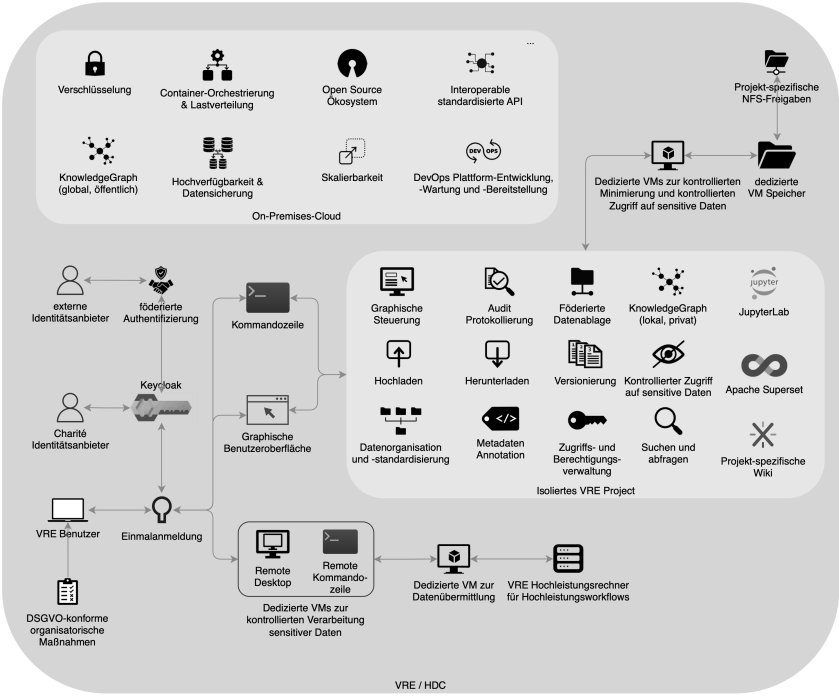
3 https://www.healthdatacloud.eu/.

4 Vorschlag für eine Verordnung über den europäischen Raum für Gesundheitsdaten, COM/2022/197 final.

5 Verordnung (EU) 2016/679 vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung), Abl. EU Nr. L 119, S. 1.

sam verarbeitet werden können. Forschungsdatensätze enthalten oft tausende von Teilnehmer:innen und Terabytes von Daten, die eine parallele Verarbeitung und große Speicherblöcke erfordern, damit Algorithmen sie effizient und rechtzeitig verarbeiten können. Darüber hinaus erfolgt der Forschungsprozess oft iterativ in einem Team, bei dem verschiedene Fachkräfte ineinander verzahnte oder unabhängige Schritte in Richtung des Projektziels liefern. Die Möglichkeit, mit dem Team in einem geschützten, gemeinsamen, skalierbaren Datei- und Betriebssystem zu arbeiten, hat mehrere Vorteile: das Team kann direkt an derselben Software und denselben Daten arbeiten, ohne umständliches und zeitraubendes Exportieren und Importieren der Daten in den beteiligten Systemen. Zu diesem Zweck bietet VRE/HDC ihren Nutzenden sowohl interaktive Prozessierung innerhalb von Containern oder Virtuellen Maschinen, aber auch sogenannte batch Prozessierung in Hochleistungsrechnern. In der VRE/HDC können Verarbeitungsabläufe mittels der gewohnten Linux und Windows graphischen und interaktiven Benutzeroberflächen entwickelt werden und nach erfolgreicher Testung auf größere Datensätze automatisch parallel auf eigens bereit gestellten Hochleistungsrechnern angewandt werden. Über grafische (GUI) und Befehlszeilen-Schnittstellen (CLI) können Forscherteams bereits bestehende oder eigens entwickelte Datenverarbeitungs-Workflows auf dedizierten virtuellen Maschinen (VM) und hochleistungsfähigen Rechnern bequem und gemeinsam ausführen. Derartig gemeinsam genutzte Systeme bringen jedoch Sicherheitsrisiken mit sich, da die verschiedenen Nutzer des Systems hauptsächlich durch Software-Isolationsmechanismen logisch getrennt sind, aber gemeinsam dieselbe Hardware benutzt wird. Bei solcher logischen Isolierung können Schwachstellen nicht ausgeschlossen werden. In der VRE/HDC ermöglicht ein rollenbasiertes Berechtigungskonzept mit abgetrennten Datenzonen die kontrollierte und überprüfbare Verarbeitung sensibler Daten über klar definierte Schnittstellen mit strenger Zugriffskontrolle. Interoperable Datenräume zielen darauf ab, rechtliche und technische Hindernisse für die gemeinsame Nutzung von Daten und die kollaborative Datenverarbeitung zu überwinden. Der Schutz der Privatsphäre sensibler Daten und die Bereitstellung von Computing-Diensten für die biomedizinische Forschung, die den nationalen und internationalen Datenschutzbestimmungen entsprechen, sind die zentralen Anwendungsfälle, für die das VRE/HDC entwickelt wurde, und diesen Zielen wurde im Entwicklungsprozess von Anfang an höchste Priorität eingeräumt. Neben Sicherheit spielen in der VRE/HDC Auffindbarkeit, Erreichbarkeit, Interoperabilität und Wiederverwendbarkeit eine große Rolle, abgekürzt durch

das Akronym FAIR (findable, accessible, interoperability and reusability). Zu diesem Zweck stellen VRE/HDC Dienste bereit, um Informationen, die die Datenherkunft betreffen, eindeutig festzuhalten, Daten zu versionieren und durch Metadaten Annotation interpretierbarer und wiederverwendbarer zu machen. Damit fungieren VRE/HDC als Referenzplattform und Orchestrierer von Rechendiensten auf gemeinsam genutzten Ressourcen die lokal, in der Cloud oder in einem hybriden Modus sicher und wiederverwendbar eingesetzt werden können. Der gesamte Quelltext und die damit verbundenen konzertierten TOMs können von anderen Forschungsgemeinschaften übernommen werden und fördern so die Entwicklung eines interoperablen Ökosystems aus Datenplattformen und Plattform-Diensten, aufgebaut von und genutzt durch eine aktive Gemeinschaft von Forschenden, wodurch Barrieren für biomedizinische Forschung und Innovation abgebaut werden.



Eine Übersicht verschiedener VRE/HDC-Funktionalitäten aus Benutzersicht ist in Abbildung 1 dargestellt. Die VRE/HDC bietet grafische Webanwendungen für den Zugriff und die Verwaltung von Ressourcen und Daten. Die Projektansicht bietet Dateiexplorer-Funktionen wie erweiterte Dateisuche, Hochladen und Herunterladen von Dateien, Metadaten Annotation durch Dateiattribute oder Metadaten-Schemata, sowie die Erstellung und Anzeige von Daten Herkunftsinformationen (Datenprovenienz). Die Projektansicht ermöglicht die Verwaltung von Nutzenden, Konfiguration von Projekteinstellungen sowie Zugang zu Verarbeitungsdiensten wie Remote Desktops und Terminals auf privaten Virtual Machines (VMs), isolierte JupyterHub container und XWiki zur erweiterten Dokumentation. Das Dateisystem eines VRE/HDC-Projektes ist in zwei Bereiche unterteilt: dem Green Room für das kontrollierte Hochladen von Daten und der Kernzone, in der die hochgeladenen Daten mit Rechenressourcen und Cloud-Diensten wie Remote Desktop und Kommandozeile auf privaten VMs verarbeitet werden können. Die separierten Datenzonen ermöglichen das kontrollierte und nachvollziehbare Hochladen, Verarbeiten und Teilen sensibler Daten. Durch Protokollierung (logging) kritischer Transaktionen und Verarbeitungsschritte kann Compliance mit der DSGVO demonstriert werden.

Die Funktionen von VRE/HDC sind durch ein Ökosystem interoperabler "Microservices" implementiert, also in Form plattformunabhängiger Container, die dynamisch mit Ressourcen ausgestattet werden können in Abhängigkeit des Nutzungsaufkommens. Dies ermöglicht, dass datenschutzrelevante Kernfunktionen wie die graphische Web-Ansicht zur Steuerung und Kontrolle der Datenflüsse, Datenzonen, projekt- und rollenbasierte Zugriffskontrollen, föderiertes Identitätsmanagement, automatische Aufnahme von Daten aus Krankenhaus-Datenquellen, sowie projektspezifische Data Warehouses für die Aufnahme und Abfrage strukturierter Datensätze und der Erfassung und Annotation von Metadaten für verschiedene Anwendungen und Infrastrukturen kombiniert und wiederverwendet werden können.

II. Datenverarbeitung in der VRE/HDC

Die Verarbeitung von Daten innerhalb der VRE/HDC verläuft ausschließlich innerhalb von Containern und VMs, die durch Zugangskontrolle geschützt sind, was es Forschenden ermöglicht, eigene, auf ihre spezifischen

Forschungsanforderungen zugeschnittene Workflows, zu erstellen und zu benutzen. VMs ermöglichen es dem Team aus Verantwortlichen und Auftragsverarbeitenden sensible Daten gemeinsam und vertraulich in einem eigenen isolierten Dateisystem zu verarbeiten. Jedes VRE/HDC-Projekt hat eigene isolierte Instanzen verschiedener Workbench-Tools einschließlich der genannten VMs, aber auch JupyterHub, Apache SuperSet und XWiki. Über Remote-Desktop- und Kommandozeilenschnittstellen können Teams effizient auf die verfügbaren Rechen- und Speicherressourcen zugreifen und so die Arbeitsabläufe flexibel an die Anforderungen des Projekts anpassen. JupyterHub ist ein Workbench Tool für die interaktive Entwicklung und Prozessierung. Nutzende können einfach Code implementieren und testen (Python, R, Julia) oder interaktive Notebooks erstellen, in denen Live-Code mit wissenschaftlichen Analysen, Abbildungen und Text integriert wird, um komplexe Analyse-Workflows zu dokumentieren. Jede:r Nutzer:in hat eine eigene private JupyterHub-Instanz, die nach eigenen Bedürfnissen angepasst werden kann, indem zum Beispiel weitere Python-Pakete installiert werden. Darüber hinaus können die Nutzer ein Terminal starten, um direkt in ihrem JupyterHub Container zu arbeiten und auf die CLI von VRE/HDC zuzugreifen, die für den Datentransfer zwischen der Kernzone des Projekts und dem JupyterHub-Container verwendet werden kann. Apache Superset ist ein Business-Intelligence-Tool zur Untersuchung und Visualisierung großer Datensätze bis in den Petabyte-Bereich. Kernzonen Daten können direkt geöffnet werden, und zusammenfassende Statistiken können erstellt und mit konfigurierbaren Dashboards visualisiert werden. VRE/HDC bietet eine vielseitige Befehlszeilenschnittstelle, genannt VRECLI, mit der Nutzende ihre Verarbeitung programmatisch steuern und automatisieren können. VRECLI kann beliebig mit den typischen Funktionalitäten eines Linux-Terminals kombiniert werden und bietet Funktionen, die zur Unterstützung kollaborativer wissenschaftlicher Verarbeitung entwickelt wurden. VRECLI ist eine ausführbare Binärdatei, die über das Guacamole VM-Terminal und das JupyterHub-Terminal verfügbar ist, aber auch für die lokale Nutzung heruntergeladen werden kann. VRECLI Benutzerbefehle ermöglichen das An- und Abmelden bei der Plattform. Projektbefehle listen die Projekte auf, auf Nutzende zugreifen können. Dateibefehle ermöglichen das Auflisten und Exportieren von Attributvorlagen (Metadatenschemata) eines Projekts und sie ermöglichen auch die Auflistung der Dateien und Ordner eines Projekts und das Hoch- und Herunterladen von Dateien oder Ordnern.

III. FAIRness

Neben dem Schutz der Privatsphäre ist das zweite Hauptanliegen des VRE/HDC die FAIR-ness: Daten auffindbar, zugänglich, interoperabel und wiederverwendbar zu machen. Die FAIR-Datengrundsätze[6] sind ein weithin akzeptierter Satz von Leitprinzipien für wissenschaftliches Datenmanagement, die Datenproduzenten und -herausgeber dazu auffordern, die maximale Nutzbarkeit von Forschungsdaten zu ermöglichen. Die VRE/HDC unterstützt Forschende bei der Einhaltung der FAIR-Prinzipien durch Werkzeuge zum Festhalten der digitalen Provenienz, zur Standardisierung und zur Versionierung mit dem Ziel, die Reproduzierbarkeit des Forschungsergebnisses zu erhöhen. Über benutzerfreundliche Schnittstellen können Ursprung und Verarbeitungsweg der Daten nachverfolgt werden, was transparente und nachvollziehbare Forschungspraktiken ermöglicht. VRE/HDC ermöglicht die Strukturierung und Validierung von Datensätzen anhand von vordefinierten Datenformaten und deren Annotation mit Metadaten auf der Grundlage von bestehenden oder benutzerdefinierten maschinenlesbaren Metadaten-Schemata. Durch Versionsverwaltung können präzise definierte und strukturierte Datenpakete archiviert, mit Zeitstempel und der Historie der Verarbeitungsschritte versehen werden und Änderungen zwischen Versionen nachvollzogen werden.

Dateien, die auf VMs produziert oder verändert wurden, können anschließend wieder in der Kernzone mit anderen Projektdaten integriert und versioniert werden. Um die Ergebnisse besser reproduzierbar und nutzbar zu machen, können die Projektdaten während der Integration in das Dateisystem des Projekts mit Provenienzinformationen versehen werden. Dies ermöglicht jede Datei mit ihren spezifischen Quell- und Ergebnisdateien sowohl den Pipelines, die zu ihrer Erstellung verwendet wurden, zu verknüpfen, was die Rekonstruktion und Visualisierung des Verlauf der Dateientwicklung mit dem Data Lineage Graph von VRE/HDC ermöglicht. Um Daten auffindbar und wiederverwendbar zu machen, bietet VRE/HDC mehrere Methoden, um Datensätze mit Metadaten zu annotieren. VRE/HDC-Nutzende können bereits bestehende, domänenrelevante, standardisierte Metadatenschemata im JSON-Format auswählen oder per Mausklick in der Benutzeroberfläche eigene Schemata erstellen. Zusätzlich zu bereits vorgegebene Metadatenschemata können Dateien und ganze Datensätze mit einem beliebigen JSON-formatierten Metadatenschema nach Wahl der

6 Wilkinson et al., 2016.

Nutzenden annotiert werden. Metadaten-Annotationen vereinfachen die Interoperabilität der Daten, da sie kontextbezogene Informationen über die Herkunft und Transformation der Daten sowie deren interne Beziehungen enthalten. Existierende Metadatensäte können leicht in den EBRAINS Knowledge Graph importiert werden, um die Sichtbarkeit, Auffindbarkeit und Wiederverwendung zu erhöhen.

IV. Informationssicherheit und Compliance

Ein wesentliches Merkmal von VRE/HDC zur Unterstützung von Informationssicherheit und Compliance ist, dass Verantwortliche und Auftragsverarbeitende sensible Daten mittels skalierbarer On-Premise Cloud Software gemeinsam, aber dennoch geschützt vor externen Zugriff, in VMs verarbeiten können. Die Hardware auf der die sensiblen Daten verarbeitet werden ist dabei durch die Virtualisierung der VMs und durch die Fernsteuerung mittels graphischen Benutzeroberflächen und Kommandozeilen-Eingaben vor direktem Zugriff geschützt. Alle datenschutzrelevanten Prozessschritte, wie das Hochladen, Herunterladen oder Teilen von Daten, werden mit Anmeldename und Zeitstempel protokolliert, um Compliance nachzuweisen. Die VRE/HDC Plattform verhindert den direkten Zugriff auf darunter liegende Ressourcen wie Speicher, Prozessoren, Datenbanken oder andere Infrastruktur-Dienste. Nutzende können nur über das grafische Webportal oder die Befehlszeilen-Tools mit Daten oder anderen Plattforminhalten interagieren. Die VMs und ihre Schnittstellen geben den Projektmitgliedern einen virtuellen Computer, auf dem sie eigene Software installieren und eigene wissenschaftlichen Arbeitsabläufe ausführen können. Die VMs können von allen Teammitgliedern gemeinschaftlich über ein persönliches Linux- oder Windows-Nutzerkonto genutzt werden. Software kann global in der VM genutzt werden und Daten können direkt im internen Dateisystem gemeinschaftlich verarbeitet werden, was manuelle Austauschvorgänge erübrigt. Die VRE/HDC wird in einer demilitarisierten Zone hinter der Firewall der Charité gehostet. Der ein- und ausgehende Datenverkehr der VMs wird durch IP-Filterung eingeschränkt: nur Verbindungen mit vorab genehmigten IP-Adressen, die zum VRE/HDC gehören, sind erlaubt, alle anderen werden abgelehnt. VMs können auch auf einzelne Nutzergruppen und die Green Room Zone beschränkt werden, um Daten, die von außerhalb importiert werden, entsprechend dem Projektziel zu minimieren bevor sie mit anderen in der Kernzone geteilt werden. Im Kubernetes Cluster

wird Isolierung zwischen den VRE/HDC Zonen und Projekten durch sogenannte Kubernetes-Namensräume und dedizierte VMs für die verschiedenen Zonen erreicht. Da die Plattformdienste mit dem Open-Source-System Kubernetes (kubernetes.io) orchestriert werden, wird Datenisolierung unter anderem mithilfe sogenannter Kubernetes-Namensräumen erreicht: zusätzlich zu Green-Room-Zone und VRE-Kernzone sorgt die Utility-Zone für die Steuerung von Datenflüsses zwischen Green Room und Kernzone. Die mit jedem Namespace verbundenen Netzwerkrichtlinien stellen sicher, dass nur autorisierter Zugriff und Datenfluss zwischen den Zonen erlaubt sind. Darüber hinaus werden die Zonen auf verschiedenen VMs gehostet und der Datenverkehr zwischen diesen VMs wird auf Netzwerkebene durch Paketfilter eingeschränkt. Die Daten jedes Projektes werden dabei in verschlüsselten Objektspeichern in dedizierten NFS-Freigaben gehalten und nur durch protokollierte Transferoperationen in der Utility-Zone gesteuert. Da der Green Room die erste Landezone für den Datenimport ist, ist die direkte Kommunikation zwischen den Diensten und der Datenfluss zwischen dem Green Room und den Kernzonen eingeschränkt und kann nur von den Projektadministratoren genehmigt werden. Die Utility-Zone enthält Backend-Dienste zur Unterstützung der Frontend-Dienste und zur Weiterleitung von Anfragen zwischen dem Green Room und der Kernzone. In der Utility Zone werden keine persönlichen Daten gespeichert, und sie kann weder auf den Green Room noch auf den VRE/HDC-Kernspeicher zugreifen. Die Dienste der Utility-Zone steuern und aggregieren den Datenfluss zwischen dem VRE/HDC und der Charité-Firewall sowie zwischen der Anwendungsprogrammierschnittstelle (API) des VRE/HDC und seinen Backend-Diensten. Darüber hinaus bietet sie Funktionen für die Verwaltung und Anpassung von Metadaten, Benachrichtigungsdienste, Projektkonfiguration und -verlauf, Benutzerkontenverwaltung und Protokollierung. Darüber hinaus werden in der Utility-Zone Kubernetes-Masterknoten bereitgestellt, die die Ausführung von Kubernetes-Pods (Gruppen von Docker-Containern mit gemeinsamem Speicher, Netzwerkressourcen und Ausführungsanweisungen) auf den Worker-Knoten verwalten und orchestrieren. Ein wesentliches sicherheitsrelevantes Merkmal der VRE/HDC ist, dass die VRE/HDC unter der EUPL 1.2 Lizenz als Open-Source Softwarepaket frei zugänglich ist und alle benutzte Software von Drittanbietern unter Open-Source Lizenzen frei im Internet zugänglich und evaluierbar ist. In Open-Source Code werden Schwachstellen durch Begutachtung durch die Entwicklergemeinschaft häufig schneller gefunden und gelöst als bei proprietärer Software.

V. Datenminimierung im Green Room

Der Green Room dient als sogenannte Staging-Area für den kontrollierten Import und die sichere Vorverarbeitung sensibler Daten. Daten können nur über den Green Room in den VRE/HDC gelangen und sind danach zunächst nur für den Uploader zugreifbar. VRE/HDC verwenden ein schrittweises Genehmigungsverfahren für die Übermittlung von Daten an die Kernzone, um zu verhindern, dass sensible Daten versehentlich an unbefugte Projektmitglieder übermittelt werden. Der Zweck des Green Rooms ist es, einen isolierten Bereich zur Verfügung zu stellen, in dem personenbezogene Daten sicher von einzelnen Nutzenden *minimiert* und vorbereitet werden können, bevor sie dem gesamten Team zur Weiterverarbeitung zugänglich gemacht werden. Nur Nutzende mit der höchsten Benutzerrolle, der Rolle "Projektadministrator", haben nach Freigabe durch den Uploader die Möglichkeit, Daten in die Kernzone zu kopieren und sie damit anderen Projektmitgliedern zur Bearbeitung zur Verfügung zu stellen. Durch die Verarbeitung von in der DSGVO definierten besonderen Kategorien von Gesundheitsdaten übernehmen diejenigen Projektmitglieder die die Zwecke und Mittel der Verarbeitung festlegen die Rolle der Verantwortlichen im Sinne von DSGVO mit den damit einhergehenden rechtlichen Verpflichtungen wie der Beachtung der Grundsätze der Zweckbindung und der Datenminimierung (Art. 5 DSGVO). Zu diesen Grundsätzen gehört beispielsweise, dass alle Informationen, die für die Zwecke der Verarbeitung nicht benötigt werden - und insbesondere direkt identifizierende Informationen wie Namen, Adressen oder Geburtsdaten - entfernt werden müssen, bevor die Daten in die Kernzone kopiert werden. Der Zugang zum Green Room ist so begrenzt, dass nur die Nutzenden, die die Daten hochgeladen haben, sowie Projektadministratoren, auf die Daten zugreifen können. Der dem Green Room zugewiesene dedizierte Speicher ist von anderen VRE/HDC-Zonen isoliert, so dass sichergestellt ist, dass nur autorisierte Dienste, die im Green Room selbst eingesetzt werden, auf die Green Room-Daten zugreifen können, nicht aber Dienste, die in anderen Zonen oder in anderen IT-Umgebungen der Charité eingesetzt werden. Es liegt in der Verantwortung der Verantwortlichen, die Daten im Green Room Zone entsprechend ihrem Zweck zu minimieren, bevor sie dem gesamten Projektteam in der VRE/HDC-Kernzone zugänglich gemacht werden. Der Green Room dient als isolierter Bereich, in dem personenbezogene Daten auf die kleinste Teilmenge reduziert (minimiert)

werden, die für die Durchführungsziele erforderlich sind. Dies ermöglicht Verantwortlichen und Auftragsverarbeitenden, dem Grundsatz der Datenminimierung (Art. 5 Abs. 1 Buchstabe c, DSGVO) Rechnung zu tragen. Es besteht die Möglichkeit für benutzerdefinierte automatisierte Workflows zur Entfernung sensibler Datenelemente und interaktive Workbench-Tools zur manuellen Überprüfung und Änderung sensibler Daten. VRE/HDC Projektadministratoren können auch Arbeitsabläufe für den automatisierten import vom Green Room in die Kernzone einrichten, sofern eine Datenminimierung automatisierbar oder nicht erforderlich ist. Beispielsweise können sensible Metadatenfelder automatisiert aus standardisiert strukturierten Daten entfernt werden. Bei nicht standardisierten Daten können Heuristiken eingesetzt werden, wie beispielsweise sogenannte "De-Facing" Algorithmen im Kontext radiologischer Bildgebung, die mittels einer Heuristik Gesichter aus MRT-Daten entfernen. Hier ist jedoch zu beachten, dass obwohl derartige Heuristiken oft sehr robust arbeiten, nichtsdestotrotz selten auftretende Fehler eine manuelle Nachkontrolle erforderlich machen. Für den automatischen Datenimport gelten die gleichen Regeln wie für den manuellen Import: Die Projektadministratoren in ihrer Rolle als Verantwortliche im Sinne der DSGVO sind dafür verantwortlich, dass alle Mitglieder des Projekts, die mit den importierten Daten in Berührung kommen, über die entsprechende Rechtsgrundlage verfügen. Die für die Datenverarbeitung Verantwortlichen und Auftragsverarbeitende sind verpflichtet, identifizierende Informationen insofern es der Verarbeitungzweck erlaubt, zu entfernen. Ein wesentlicher Aspekt hierbei ist, dass selbst nach erfolgter Minimierung Gesundheitsdaten, die mit einer natürlichen, lebenden Person in Verbindung gebracht werden können, weiterhin durch die Datenschutzgesetze geschützt sind, da sie eine besondere Kategorie persönlicher Daten darstellen, die Rückschlüsse auf den Gesundheitszustand des Individuums zulassen. Selbst nach der Entfernung von direkt identifizierenden Informationen wie Patientennamen, Gesichtern, Adressen oder Geburtsdaten (sogenannte Pseudonymisierung) gelten die Daten immer noch als personenbezogene Daten im Sinne der DSGVO, wenn sie durch die Verwendung zusätzlicher Informationen einer natürlichen Person zugeordnet werden könnten (Erwägungsgrund 26 DSGVO) und somit die Schutzanforderungen dennoch gelten. Typische biomedizinische Daten wie radiologische Bilddaten oder genetisches Material enthalten umfangreiche gesundheitsbezogene Informationen und sind deshalb mit dem gleichen Schutz zu behandeln, wie Daten die direkt identifizierende Merk-

male enthalten. Eindrucksvolle Beispiele von simulierten und konkreten Datenlecks verdeutlichen wie Daten obwohl pseudonymisiert, später durch Zusammenführung mit anderen Daten zur Re-identifizierung der Teilnehmenden genutzt werden konnten.[7] Beispielsweise konnten Hacker kürzlich die Stammbaumdaten von 6,9 Millionen Menschen erbeuten und Teile der Daten wurden anschließend auf einem Online Schwarzmarkt angeboten mit dem Ziel die ethnische Herkunft der Personen zu identifizieren.[8] Der Green Room bietet daher eine isolierte Umgebung, in der Daten sicher überprüft und sensible Informationen entfernt werden können, bevor die Daten in die Kernzone übertragen werden.

VI. Zugangskontrolle

Ein wesentlicher Grundsatz der DSGVO ist sicherzustellen, dass personenbezogene Daten durch geeignete TOMs vor unrechtmäßigem Zugriff und unrechtmäßiger Verarbeitung geschützt werden, einschließlich dem Schutz vor unbeabsichtigtem Verlust oder unbeabsichtigter Schädigung. Im Kontext von Online-Plattformen betrifft dies neben Maßnahmen zur Verhinderung des Zugriffs auf die physischen Datenverarbeitungssysteme (wie beispielsweise Gebäudesicherung) vor allem die logische Zugangskontrolle, also Maßnahmen zur Verhinderung des unbefugten Zugriffs auf bestimmte Dateisysteme oder Dienste der Software-Plattform. Der Zugang zur VRE/HDC wird durch eine ineinander verzahnte rollen- und projektbasierte Zugangskontrolle verwaltet und durch transparente Nutzungs- und Zugangsrichtlinien geregelt um eine DSGVO-konforme Steuerung und Benutzung der VRE/HDC zu gewährleisten. Die Aufnahme eines neuen Projekts erfolgt nach einem standardisierten Verfahren: Forschende können den Zugang beantragen, indem sie eine Datenschutzfolgeabschätzung (DSFA, Art. 35 DSGVO) der geplanten Verarbeitungstätigkeit erstellen und diese von allen beteiligten Verantwortlichen, Auftragsverarbeitenden und Datenschutzbeauftragten akzeptiert wird. Sobald ein Projekt ein posi-

7 Gymrek, Melissa, et al. "Identifying personal genomes by surname inference." Science 339.6117 (2013): 321-324; Rocher, Luc, Julien M. Hendrickx, and Yves-Alexandre De Montjoye. "Estimating the success of re-identifications in incomplete datasets using generative models." Nature communications 10.1 (2019): 1-9.

8 https://www.spiegel.de/netzwelt/web/hacker-erbeuten-stammbaumdaten-von-6-9-millionen-menschen-a-b34b1a4d-779e-4717-9583-dbdf9335bd10; https://www.sec.gov/ix?doc=/Archives/edgar/data/1804591/000119312523287449/d242666d8ka.htm.

tives Votum der beteiligten Datenschutzbeauftragten erhalten hat, erhält zunächst der Projektadministrator Zugang zu dem neu angelegten Projekt. Der Projektadministrator kann dann weitere Nutzer mit verschiedenen Befugnissen zum Zugriff auf die Projektdaten einladen. Die zusätzlichen Rollen „Projektmitarbeiter" bzw. „Projektmitwirkende" ermöglichen eine fein abgestufte Zugriffskontrolle, um den Umfang der Daten, die verschiedenen Nutzenden zugänglich sind, zu minimieren. So kann ein Projektmitarbeiter zwar Daten verarbeiten, die von einem Projektadministrator in der Kernzone bereitgestellt wurden, er kann jedoch keine Dateien aus dem Green Room in die Kernzone kopieren oder Green-Room-Dateien durchsuchen oder herunterladen, die von einem anderen Projektmitglied bereitgestellt wurden. In ähnlicher Weise kann ein Projektmitarbeiter Daten in den Green Room hochladen, aber er kann keine Daten in der Kernzone sehen oder darauf zugreifen. VRE/HDC-Nutzende können Mitglieder mehrerer Projekte sein und in jedem Projekt unterschiedliche Rollen einnehmen, um ein angemessenes Maß an Zugang zu ermöglichen, das den tatsächlichen Anforderungen und rechtlichen Grundlagen der Verarbeitung innerhalb des jeweiligen Projekts entspricht. Die Anmeldung an die VRE/HDC erfordert ein Nutzerkonto innerhalb des Charité Identitätsmanagementsystems, was einer formalisierten Antragsprozedur inklusive der Nennung der letzten sechs Ziffern des Personalausweises bedarf. Das System erfordert sichere Passwörter, die alle 120 Tage geändert werden müssen, um weiterhin Zugang zu erhalten. Die Passwörter werden zentral vom Charité Identitätsmanagementsystem verwaltet. Die Anforderungen an die Komplexität der Passwörter und die Zeitüberschreitung bei Inaktivität reduzieren das Risiko eines unbefugten Zugriffs. Die Identität der VRE/HDC-Nutzenden wird zwischen dem VRE/HDC-Identitäts- und Zugriffsmanagementsystem Keycloak und dem Charité System Microsoft Active Directory zusammengeführt wobei die Authentifizierung durch das Charité System durchgeführt wird und Passwörter nicht im VRE/HDC selbst gespeichert werden. Keycloak wird auch für die Authentifizierung der Kommunikation zwischen dem VRE/HDC-Frontend, dem API-Gateway (das alle Backend-Dienste verbindet) und den Workbench-Tools verwendet, basierend auf dem OpenID Connect Authentifizierungsprotokoll, über das die Dienste digital signierte Zugriffstoken austauschen, um die Authentizität und Zulässigkeit von Anfragen festzustellen. Keycloak bietet auch Single-Sign-On (SSO)-Funktionen für Anwendungen von Drittanbietern, die über das VRE/HDC-Portal bereitgestellt werden (z. B. JupyterHub, XWiki). Charité-Beschäftigte mit bestehendem Nutzerkonto können direkt

dem VRE/HDC Testprojekt hinzugefügt werden, während externe Nutzer zunächst ein Registrierungsverfahren durchlaufen und sich mit einem Personaldokument und dem Nachweis eines Vertragsverhältnisses mit der Charité ausweisen müssen, um ein Charité Nutzerkonto zu erhalten.

VII. Verschlüsselung

Bei der Verschlüsselung werden Informationen (Texte, Dateien, usw.) in einen Code umgewandelt, der ohne den richtigen Schlüssel nicht interpretiert werden kann. Dadurch werden Risiken, die durch unerwartete Datenlecks entstehen können verringert, da verschlüsselte Inhalte für Dritte, die nicht im Besitz des richtigen Schlüssels sind, grundsätzlich unlesbar sind, wodurch gespeicherte oder übertragene Daten geschützt werden. Daten, die in die VRE/HDC-Plattform importiert oder exportiert oder innerhalb der VRE/HDC transferiert werden ("data in flight"), sind während der Übertragung verschlüsselt. Darüber hinaus werden Daten, die in der VRE/HDC gespeichert werden ("Daten im Ruhezustand"), durch Hardwareverschlüsselung der Speichermedien und durch Speicherung der sensiblen Daten als verschlüsselte Objekte im VRE/HDC Objektspeicher geschützt. Diese Maßnahmen dienen dazu, das Risiko einer unbefugten Offenlegung zu minimieren, falls die Daten während der Übertragung abgefangen werden oder die Speichermedien physisch aus der IT-Umgebung entfernt werden. Für die Verarbeitung der Daten müssen diese in der Regel entschlüsselt werden und es werden daher in der VRE/HDC für die eigentliche Verarbeitung zusätzliche Sicherheitsmechanismen wie die isolierte Verarbeitung in privaten VMs oder Containern eingesetzt.

VIII. Rechenzentrum

Das VRE/HDC nutzt die generische Big-Data-Infrastruktur einschließlich Rechen- und Speicherressourcen des Geschäftsbereiches IT der Charité Universitätsmedizin Berlin – einer vom Bundesamt für Sicherheit in der Informationstechnik (BSI) zertifizierten kritischen Infrastruktur in Deutschland.

IX. Ethischer, gesellschaftlicher und rechtlicher Governance-Rahmen

Da der Personenbezug oft nicht aus den Gesundheitsdaten entfernt werden kann und dieser das ausdrückliche Ziel der personalisierten Medizin ist, muss die Verarbeitung durch "Datenschutz durch Technik und durch datenschutzfreundliche Voreinstellungen" geschützt werden, um sicherzustellen, dass personenbezogene Daten nicht einer unbestimmten Anzahl natürlicher Personen zugänglich gemacht werden (Art. 25 DSGVO). Angesichts der Notwendigkeit für Datensicherheit und Datenschutz legt VRE/HDC größten Wert auf die Gewährleistung der Vertraulichkeit, Integrität und Verfügbarkeit von Forschungsdaten. Die Plattform hält sich an die Grundsätze der DSGVO und umfasst robuste Verschlüsselungsalgorithmen, Zugriffskontrollen und Anonymisierungstechniken zum Schutz sensibler Daten. Datenschutzvorschriften wie die DSGVO und das britische Datenschutzgesetz stellen bestimmte Anforderungen an den Umgang mit personenbezogenen Daten. Das Hauptziel solcher Rechtsvorschriften besteht darin, dass der Datenschutz bereits in die Gestaltung eines Verarbeitungssystems "eingebaut" wird, so dass personenbezogene Daten zu keinem Zeitpunkt (standardmäßig) einer unbestimmten Anzahl natürlicher Personen zugänglich gemacht werden. Weiterhin regelt die DSGVO, dass die in den Daten abgebildete Person (die betroffene Person) jederzeit die letztendliche Kontrolle über ihre Daten behält. So kann die betroffene Person beispielsweise jederzeit ihre Zustimmung zur Verwendung ihrer personenbezogenen Daten zurückziehen und verlangen, dass alle Kopien oder Replikationen dieser personenbezogenen Daten unverzüglich gelöscht werden. Das Gesetz stellt klar, dass die für die Datenverarbeitung Verantwortlichen (jede Person die über die Zwecke und Mittel der Verarbeitung entscheidet, also beispielsweise eine Professorin die einen Doktorstudenten beauftragt eine Analyse durchzuführen) die Verantwortung dafür tragen, dass sie im Einklang mit dem Gesetz handeln. Hierbei ist es wichtig es zu beachten, dass es in Computersystemen ohne angemessene TOMs üblicherweise unmöglich ist, die Kontrolle über digitale Objekte zu behalten, sobald sie einmal anderen zur Verfügung gestellt wurden. Auch andere Rechte betroffener Personen sind mit einem enormen bürokratischen Aufwand Seitens der Forschenden verbunden, wenn keine angemessene Infrastruktur vorhanden ist die es erlaubt Daten kontrolliert und protokolliert zu verarbeiten. So haben die betroffenen Personen unter anderem das Recht, die Namen und Kontaktdaten aller für die Verarbeitung Verantwortlichen zu erfahren, als auch die Empfänger der personenbezogenen

Daten, den Zeitraum, für den die Daten gespeichert werden, und sogar den Zugang zu diesen Daten zu erhalten. Die DSGVO schreibt daher vor, dass Datenschutz in die Architektur der Verarbeitungstätigkeit eingebettet wird, indem geeignete TOMs getroffen werden, um die Daten jederzeit zu schützen, unabhängig davon, ob sie aktiv verarbeitet werden oder ruhen, und die Verarbeitungstätigkeit im Hinblick auf diese Maßnahmen laufend zu protokollieren. Organisatorische Maßnahmen beziehen sich beispielsweise auf die Datenminimierung und die Begrenzung der Menge personenbezogener Daten auf das für die Verarbeitung erforderliche Maß sowie auf die Löschung von Daten, die nicht mehr benötigt werden. Zu den technischen Maßnahmen gehören nicht nur Maßnahmen zum Schutz der laufenden Vertraulichkeit, wie die Verschlüsselung. Wichtig ist, dass auch Maßnahmen zum Schutz der Integrität, Verfügbarkeit und Widerstandsfähigkeit der Verarbeitungssysteme und -dienste gesetzlich vorgeschrieben sind (Art. 5 und 32 DSGVO). Dazu gehört die Fähigkeit, die Verfügbarkeit und den Zugang zu personenbezogenen Daten im Falle eines physischen oder technischen Vorfalls zeitnah wiederherzustellen, sowie ein Verfahren zum regelmäßigen Nachweis der Wirksamkeit der getroffenen Maßnahmen. Personen, die infolge eines Verstoßes gegen diese Vorschriften einen Schaden erlitten haben, haben das Recht, von den Verantwortlichen oder den Auftragsverarbeitenden Schadenersatz zu erhalten, was zu Geldstrafen von bis zu 20 Millionen Euro oder 4 % des weltweiten Jahresumsatzes des Unternehmens führen kann. Bei mehreren Verantwortlichen und Auftragsverarbeitenden haften zunächst all Verantwortlichen und Auftragsverarbeitenden für den gesamten Schaden, damit ein wirksamer Schadensersatz für die betroffene Person sichergestellt ist (Art. 82 DSGVO). Aus diesem Grund ist es wichtig, dass die Verantwortlichen in der Lage sind, die für die Verarbeitung verwendeten Drittdienste sorgfältig zu überprüfen. Um Schwachstellen zu beseitigen und die Einhaltung der Vorschriften nachzuweisen, hat der VRE/HDC seinen gesamten Quellcode und seine Dokumentation als Open Source (github.com/vre-charite) veröffentlicht und für jedermann zugänglich gemacht. Über ihre technischen Fähigkeiten hinaus fungiert die VRE/HDC-Plattform als Referenzimplementierung für die Forschungsgemeinschaft und verkörpert die besten Praktiken und Standards, die für die Einhaltung der DSGVO und die Erfüllung der EHDS-Anforderungen erforderlich sind. Als Open-Source-Lösung fördert sie die Zusammenarbeit und Mitwirkung und unterstützt so ein lebendiges Ökosystem von Forschern, Entwicklern und Interessenvertretern, die sich für die Förderung wissenschaftlicher Entdeckungen einsetzen und gleichzeitig

strenge Vorschriften zum Schutz der Privatsphäre und des Datenschutzes einhalten.

X. Rechtmäßige Grundlage

Die DSGVO verlangt von den Verantwortlichen, dass sie personenbezogene Daten nur auf Basis einer spezifischen Auswahl von Rechtsgrundlagen erheben, speichern und verwenden. Verantwortliche können Gesundheitsdaten im VRE/HDC nur dann verarbeiten, wenn sie eine rechtmäßige Grundlage nach Art. 6 und 9 (DSGVO) für die Verarbeitung besonderer Kategorien personenbezogener Daten besitzen und den Risiken bei der Verarbeitung dieser besonderen Datenkategorie (Daten die Auskunft über die Gesundheit lebender Personen geben) adäquat Rechnung tragen. Im Falle solcher besonderen Datenkategorien ist es notwendig, sowohl eine Rechtsgrundlage für die allgemeine Verarbeitung als auch eine zusätzliche, separate Bedingung für die Verarbeitung dieser besonderen Datenkategorie zu bestimmen (Art. 9 DSGVO). Daten, die sich auf die Gesundheit von Personen beziehen, sind eine besondere Kategorie von Daten (Art. 4 Abs. 15 DSGVO), deren Verarbeitung grundsätzlich verboten ist (Art. 9 Abs. 1 DSGVO), es sei denn, die in den Daten dargestellte Person (die betroffene Person) hat ausdrücklich in die Verarbeitung für einen oder mehrere festgelegte Zwecke eingewilligt (Art. 9 Abs. 2 Buchstabe a DSGVO) oder die Verarbeitung ist für wissenschaftliche Zwecke erforderlich (Art. 9 Abs. 2 Buchstabe j DSGVO), unter der Bedingung, dass das wesentliche Recht auf Datenschutz unter Verwendung von Garantien, die die Einhaltung des Grundsatzes der Datenminimierung gewährleisten, gewahrt wird (Art. 89 Abs. 1 DSGVO). Für eine rechtmäßige Verarbeitung von Gesundheitsdaten im VRE/HDC zu Forschungszwecken ist es daher erforderlich, dass die betroffene Person in die Verarbeitung für einen oder mehrere bestimmte Zwecke eingewilligt hat (Art. 6 Abs. 1 Buchstabe a)) und dass sie ausdrücklich in die Verarbeitung von Gesundheitsdaten für einen oder mehrere festgelegte Zwecke eingewilligt hat (Art. 9 Abs. 2 Buchstabe a)) und dass der zuständige lokale institutionelle Ethikrat die Verarbeitung genehmigt hat. Die Einwilligung als Rechtsgrundlage bringt bestimmte zusätzliche Verpflichtungen mit sich: Die Einwilligung muss freiwillig, ausdrücklich, in Kenntnis der Sachlage und unmissverständlich erteilt werden, und sie muss jederzeit widerrufen werden können. Die Verantwortlichen müssen dabei sicherstellen, dass der Widerruf der Einwilligung genau so

einfach ist wie die Erteilung der Einwilligung: es muss für die betroffenen Personen genauso einfach sein, ihre Einwilligung zu widerrufen, wie es für die Verantwortlichen war, die Einwilligung einzuholen. Die ausdrückliche Einwilligung für einen oder mehrere festgelegte Zwecke voraus, dass das Ersuchen um Einwilligung in verständlicher und leicht zugänglicher Form in einer klaren und einfachen Sprache erfolgt und von anderen Sachverhalten klar zu unterscheiden ist. Die ausführliche Dokumentation und die Datenschutzrichtlinien von VRE/HDC helfen dabei, diese Informationen an die betroffenen Personen weiterzugeben und sie adäquat zu informieren. In der VRE/HDC wird daher die rechtmäßige Grundlage vor Beginn der Verarbeitung im Rahmen der DSFA ermittelt, noch bevor ein VRE/HDC-Projekt erstellt wird.

XI. Datenschutz-Folgenabschätzung (Art. 35 DSGVO)

Wenn die Verarbeitung wahrscheinlich ein hohes Risiko für die Rechte und Freiheiten natürlicher Personen mit sich bringt, muss der für die Verarbeitung Verantwortliche eine Abschätzung der Auswirkungen der geplanten Verarbeitungen auf den Schutz personenbezogener Daten vornehmen. Um die Rechtmäßigkeit der Verarbeitung zu gewährleisten, ist daher das Hochladen von personenbezogenen Daten in das VRE/HDC erst nach Durchführung einer DSFA zulässig. Der VRE/HDC unterstützt seine Nutzer durch die Bereitstellung von DSFA-Vorlagen für die Verarbeitung von Gesundheitsdaten. Der Zweck der DSFA ist hierbei die Einhaltung der DSGVO bei der geplanten Verarbeitungstätigkeit nachzuweisen, indem die Zwecke der Verarbeitung, die Art der verarbeiteten Daten, die Zugriffsberechtigten, die Maßnahmen zum Schutz der Daten und der Zeitpunkt der geplanten Löschung angegeben werden. Die DSGVO schreibt die Durchführung einer DSFA vor, wenn die Verarbeitung "voraussichtlich ein hohes Risiko für die Rechte und Freiheiten natürlicher Personen zur Folge" hat (Art. 35). Die Aufnahme eines neuen Projekts im VRE/HDC erfolgt daher erst nachdem die Verantwortlichen eine DSFA der geplanten Verarbeitungstätigkeit erstellt haben und diese von allen beteiligten Verantwortlichen, Auftragsverarbeitenden und Datenschutzbeauftragten akzeptiert wurde. Sollte aus der DSFA hervorgehen, dass die Verarbeitung ein hohes Risiko zur Folge hätte, sofern die Verantwortlichen keine Maßnahmen zur Eindämmung des Risikos treffen (einschließlich der TOMs die durch die VRE/HDC bereitgestellt werden) ist es die Pflicht der Verantwortlichen vor

der Verarbeitung die entsprechenden Aufsichtsbehörden zu konsultieren (Art. 36 DSGVO). Die VRE/HDC stellt Verantwortlichen eine detaillierte DSFA-Vorlage zur Verfügung, um diesen Prozess zu erleichtern. Die DSFA-Vorlage enthält allgemeine Verarbeitungsschritte der VRE/HDC für einen typischen Anwendungsfall zur Verarbeitung von Gesundheitsdaten und legt alle TOMs zur Risikominderung explizit dar, um die Anfertigung der DSFA für das Gesamtprojekt, als auch die Kontrolle durch Datenschutzbeauftrage, zu erleichtern. Da die wesentlichen Risiken, Verarbeitungsschritte und TOMs für typische Anwendungsfälle zur Verarbeitung von Gesundheitsdaten vergleichbar sind, kann der DSFA-Prozess für VRE/HDC Projekte erleichtert werden, so dass Verantwortliche lediglich risikorelevante Änderungen mit den Datenschutzbeauftragten diskutieren müssen. Sobald ein Projekt ein positives Votum der beteiligten Datenschutzbeauftragten erhalten hat, erhält zunächst ein/e Verantwortliche/r Zugang als Projektadministrator zu dem neu angelegten Projekt und kann anschließend die spezifizierten weiteren Verantwortlichen oder Auftragsverarbeitenden hinzufügen um die geplante Verarbeitungstätigkeit zu beginnen.

XII. Rechenschaftspflicht und geteilte Verantwortung

Als organisatorische Maßnahme unterscheidet die VRE/HDC spezifische Rollen für verschiedene Nutzerkategorien deren Berechtigungen sich auf die Verantwortlichkeiten der in der DSGVO definierten Rollen beziehen. Die für die Verarbeitung Verantwortlichen werden in der DSGVO definiert als jede Person, Behörde, Einrichtung oder andere Stelle, die allein oder gemeinsam mit anderen über die Zwecke und Mittel der Verarbeitung personenbezogener Daten entscheidet (Art. 4 Abs. 7 DSGVO), während die Auftragsverarbeitenden diese Daten lediglich im Auftrag und auf Anweisung der Verantwortlichen verarbeiten (Art. 4 Abs. 8 DSGVO). Die Verantwortlichen sind rechtlich verpflichtet, die Privatsphäre der personenbezogenen Gesundheitsdaten der betroffenen Personen zu schützen (Art. 24 DSGVO). Der Hauptzweck des VRE/HDC besteht darin, Verantwortliche und Auftragsverarbeitende dabei zu helfen, nachzuweisen, dass die Verarbeitung auf rechtmäßige Weise erfolgt, indem sie die beschriebenen technischen und organisatorischen Maßnahmen für den Datenschutz und die Einhaltung der Rechtsvorschriften umsetzen. Erst nachdem die DSFA und ein Datenverarbeitungsvertrag basierend auf den Standardvertragsklauseln der Europäischen Kommission abgeschlossen wurden und

weitere gesetzlich vorgeschriebene Dokumente vorgelegt wurden (z.B. eine Ethikgenehmigung von der zuständigen Ethikkomission entsprechend dem Humanforschungsgesetzes in Verbindung mit den entsprechenden Landesgesetzen), wird von einem Plattformadministrator ein spezifisches VRE/HDC-Projekt erstellt und die gelisteten Verantwortlichen und Auftragsverarbeitenden dem Projekt hinzugefügt. Die für die Verarbeitung Verantwortlichen können dann den in der DSFA spezifizierten Gesundheitsdatensatz importieren, um ihn kontrolliert allen Teammitgliedern für die Erfüllung der beabsichtigten Zwecke der Verarbeitung zur Verfügung zu stellen beziehungsweise ihn vorher entsprechend des Zweckes zu minimieren.

XIII. Datenverarbeitungsvertrag

Die Verarbeitung durch einen Auftragsverarbeiter erfolgt auf der Grundlage eines Vertrags, der den Auftragsverarbeiter in Bezug auf den Verantwortlichen bindet und in dem Gegenstand, Dauer, Art und Zweck der Verarbeitung, die Art der personenbezogenen Daten, die Kategorien betroffener Personen und die Pflichten und Rechte des Verantwortlichen festgelegt sind (Art. 28 DSGVO). Der Vertrag enthält unter anderem die gegenseitige Zusicherung der Beteiligten die Bestimmungen der DSGVO einzuhalten, was einschließt die personenbezogenen Daten zu schützen und nachweisen zu können, dass die jeweiligen Anforderungen in Bezug auf die Verarbeitung und die rechtliche Verantwortung erfüllt sind. VRE/HDC bietet eine Vorlage für Datenschutzvereinbarungen anhand von Standardvertragsklauseln, die von der Europäischen Kommission publiziert wurden und die Vorgaben der DSGVO ausdrücklich niederlegen. Der Datenverarbeitungsvertrag klärt die gemeinsame Verantwortung der Beteiligten und die technischen und organisatorischen Maßnahmen zum Schutz der Verarbeitung. Mit dem Datenverarbeitungsvertrag schließen Verantwortliche für Gesundheitsdaten einen Vertrag mit der Charité – Universitätsmedizin Berlin als Auftragsverarbeiterin und Betreiberin der VRE/HDC. Das Dokument ermöglicht es den Beteiligten Vereinbarungen hinsichtlich der rechtlichen Konformität, Sicherheit und Integrität der Daten, Haftung, Schadensersatz und Durchsetzbarkeit vertraglich festzuhalten und zu überprüfen.

XIV. HPC

Zusätzlich zu den Rechenressourcen, die von den VMs bereitgestellt werden, stellt die HPC-Infrastruktur der Charité Prozessorzeit und Speicher für die Ausführung ressourcenintensiver Aufgaben zur Verfügung. Der VRE Kommandozeilen-Client ermöglicht Nutzern, ressourcenintensive Aufgaben an das Charité IT HPC zu übermitteln. Typische Parameter von HPC-Jobs (wie die Anzahl der parallelen Prozesse oder der jedem Prozess zur Verfügung stehende Speicher) können mit JSON-Dateien konfiguriert werden. Der Befehl ermöglicht auch die Abfrage des Status und Ergebnisses eines eingereichten Jobs sowie Standardfehler- und andere Standardausgaben weiterzuleiten. Darüber hinaus ermöglicht der Befehl die Abfrage von Informationen über die verfügbaren Partitionen sowie die verfügbaren Knoten und ihre Hardwarekonfiguration. Nutzende haben die Auswahl zwischen CPU und GPU Chipsätzen, sowie zwischen Knoten mit hohen Speicheranforderungen oder langer Laufzeit. Der HPC-Dienst ist für die Ausführung von containerisierten Workflows konzipiert, die keine Interaktion erfordern und automatisch ausgeführt werden können, im Gegensatz zu den VMs und JupyterHub, die in erster Linie für interaktive Berechnungen und weniger ressourcenintensive Aufgaben verwendet werden. Benutzer können sich über den VRE Kommandozeilen-Client (CLI) auf einem Projekt-spezifischen remote Desktop VM mit dem HPC-System verbinden, Daten austauschen, Jobs konfigurieren, starten und die Ergebnisse abholen. Die CLI bietet Befehle zur Authentifizierung, zur Übertragung von Daten und Software, zur Konfiguration und Übermittlung von Aufträgen, zur Abfrage des Status eines Auftrags und zur Integration der Ergebnisse zurück in das Projekt. HPC-Aufträge können im JSON-Format konfiguriert werden, wobei Auftragsparameter wie die Anzahl der parallelen Aufgaben, die Namen der HPC-Knoten für die Ausführung, der Speicher pro CPU, die Nutzung der GPU-Ressourcen, Umgebungsvariablen oder E-Mail-Adressen für den Empfang von Benachrichtigungen angegeben werden können.

XV. Versionskontrolle, Integritätsschutz und Herkunftsnachweis

Detaillierte Informationen über die Bearbeitungshistorie (digitale Provenienz) zur Nachverfolgung von Änderungen eines Datensatzes sind für die Wiederverwendung im Sinne der FAIR Prinzipien notwendig, damit Forschende verstehen können wie die Daten erzeugt wurden, in welchen

Zusammenhängen sie wiederverwendet werden können und wie zuverlässig die enthaltenen Informationen sind. VRE/HDC bieten Werkzeuge zur Erzeugung und (Wieder-)Verwendung von Metadatenschemata mittels GUI oder JSON- Dateien. Die bereitgestellten Provenance-Werkzeuge ermöglichen die Beschreibung von Forschungsprodukten mit einer beliebigen Anzahl an Attributen, um die Daten für künftige Forschung *wiederverwendbar* zu machen. Metadatenschemas können flexibel gestaltet und kombiniert werden um Eigenschaften des Datensatzes zu beschreiben, wie beispielsweise Nutzungslizenz, Verweise auf den Rohdatensatz, verwendete Verarbeitungspipelines, Informationen, die für die ordnungsgemäße Nutzung und Interpretation der im Datensatz enthaltenen Informationen und Kenntnisse erforderlich sind, wie z. B. die wissenschaftliche Definition von Variablennamen, physikalischen Einheiten oder die räumliche Lokalisierung von Datenelementen. Jede Datei kann mit anderen Dateien verknüpft werden, aus denen sie zusammen mit den durchgeführten Verarbeitungsschritten abgeleitet wurde, wodurch vollständige Abstammungslinien digitaler Objekte generiert werden können. Dateien, die durch Verarbeitungsschritte erstellt wurden können ihren Quelldateien und Verarbeitungspipelines zugeordnet werden, so dass Provenance-Informationen gespeichert werden können, aus dem die Eingaben und Prozesse hervorgehen, die zur Erstellung jeder Datei verwendet wurden. Eine Abstammungslinie kann verwendet werden, um den Verlauf jeder Datei vom Hochladen bis zu ihrem aktuellen Zustand sowie allen daraus generierten Ergebnissen anzuzeigen, was nützlich ist, um den Lebenszyklus digitaler Objekte zu verstehen und zu reproduzieren. Die Provenance-Historie des Datasets ermöglicht es, Art und Zeitpunkt sowie die bei der Verarbeitung involvierten Nutzende, eindeutig nachzuvollziehen. Die zugehörigen VRE/HDC Dataset-Werkzeuge erlauben unveränderliche "snapshots" von Datensätzen einschließlich verschiedener Versionsnummern anzufertigen und zu teilen. Mit dem zugehörigen Download-Befehl können Nutzende angeben, welche Version eines Datensatzes heruntergeladen werden soll. An einem Dataset vorgenommene Änderungen werden automatisch in Echtzeit verfolgt und können im Dataset Provenance Activity Stream angezeigt werden. Zusätzlich zur automatischen Änderungsverfolgung können Nutzende auch manuell die Erstellung eines Schnappschusses des spezifischen Status eines Datasets zu einem bestimmten Zeitpunkt auslösen. Jede neue Dataset-Version wird dann mit einer eindeutigen Versionsnummer versehen und schreibgeschützt, um sicherzustellen, dass die Daten in ihrem exakten Zustand bleiben. Überarbeitungen, die an einem Datensatz und seinen

Metadaten vorgenommen wurden, werden von VRE/HDC als zusätzliche Metadaten des Datensatzes nachverfolgt und seine Revisionsgeschichte (Provenance) gespeichert.

XVI. Schlussfolgerung

In diesem Beitrag haben wir eine neuartige Open-Source virtuelle Forschungsumgebung für die DSGVO-konforme Verarbeitung von Gesundheitsdaten vorgestellt. Durch die Nutzung von VMs, Kubernetes-Clustern und Containerisierung bietet VRE/HDC eine sichere und skalierbare Umgebung für Biomedizinische Forschung. Mit ihrem Schwerpunkt auf Datensicherheit und FAIRness und ihrem Open-Source Quellcode (https://github.com/vre-charite) fungiert VRE/HDC als Referenzimplementierung um kollaborative Forschung und Innovation zu fördern.

Finanzierungsquellen

Diese Arbeit wurde unterstützt durch die Virtuelle Forschungsumgebung an der Charité Berlin - ein Knotenpunkt der EBRAINS Health Data Cloud. Wir danken außerdem für die Unterstützung durch H2020 Research and Innovation Action Grant Human Brain Project SGA3 945539; H2020 Research and Innovation Action Grant Interactive Computing E-Infrastructure for the Human Brain Project ICEI 800858; H2020 Research and Innovation Action Grant EOSC VirtualBrainCloud 826421; Horizon Europe Research and Innovation Action Grant AISN 101057655; Horizon Europe Research Infrastructures Grant EBRAINS-PREP 101079717; European Innovation Council PHRASE 101058240; Horizon Europe Research Infrastructures Grant EBRAIN-Health 101058516; H2020 European Research Council Grant ERC BrainModes 683049; JPND ERA PerMed PatternCog 2522FSB904; Digital Europe TEF-Health 101100700; Berlin Institute of Health & Foundation Charité; Johanna Quandt Exzellenzinitiative; Deutsche Forschungsgemeinschaft SFB 1436 (Projekt-ID 425899996); Deutsche Forschungsgemeinschaft SFB 1315 (Projekt-ID 327654276); Deutsche Forschungsgemeinschaft SFB 936 (Projekt-ID 178316478)); Deutsche Forschungsgemeinschaft SFB-TRR 295 (Projekt-ID 424778381); Deutsche Forschungsgemeinschaft SPP Computational Connectomics RI 2073/6-1,

RI 2073/10-2, RI 2073/9-1; Deutsche Forschungsgemeinschaft Klinische Forschergruppe BECAUSE-Y (Projekt-ID 504745852).