Datenmanagement bei popgen

Huberta von Eller-Eberstein, Lukas Gundermann, Michael Krawczak, Stefan Schreiber, Andreas Wolf

popgen - Populationsrepräsentative Bevölkerungsstichprobe und Krankheitskohorte Nord-Schleswig-Holstein
Universitätsklinikum Schleswig-Holstein, Campus Kiel
Brunswiker Straße 12
24105 Kiel
Huberta von Eller-Eberstein <eberstein@popgen.de>
Lukas Gundermann <gundermann@datenschutzzentrum.de>

Abstract: At the core of the popgen data security concept lies the maintenance of two separated databases ("Recruitment", "Laboratory") that dissociate personal and phenotype data from genotypes. The two data types are labelled by independent identifiers (RC and PsID), which can only be connected via a dedicated, specifically protected, trustee server. All data transfer from the recruitment office to the laboratory management system and to the data analysis platform proceeds through the trustee server. This computer exclusively holds a table encoding the RC/PsID relationship. The trustee server automatically generates the second identifier, PsID, upon request. Identifier PsID is used to re-label DNA samples before producing genotypes and to re-label pertinent reduced sets of personal and phenotypic data before forwarding them to the data analysis platform.

1 Einleitung

Dieser Beitrag hat zum Ziel, anhand einer tatsächlich existierenden Biomaterialbank (BMB) zu illustrieren, in welcher Weise durch das Zusammenspiel verschieden Institutionen eine sichere Abschottung von identifizierenden Daten der Probanden erreicht werden kann. Dabei wird der Schwerpunkt auf die organisatorische Ausgestaltung, Prozesse und Datenflüsse gelegt. Deren Realisierung im Projekt popgen wird mit den Vorgaben vergleichen, die die Telematikplattform für Medizinische Forschungsnetze (TMF) definiert hat [TMF06]. Einzelheiten zu eher rechtlichen Fragen (Einwilligungserklärung, Zulassung von Forschungsvorhaben, Weitergabe von Proben an andere Einrichtungen) können hier nicht behandelt werden.

popgen ist ein Projekt im Rahmen des Nationalen Genomforschungsnetzes (NGFN) mit dem Ziel, eine umfassende Biomaterialdatenbank aufzubauen, um die Zusammenhänge von Krankheiten und genetischen Merkmalen zu untersuchen. Bei popgen werden dafür die administrativen Grundlagen geschaffen, jedoch keine eigenen Forschungsvorhaben verfolgt. Einzelne medizinisch-wissenschaftliche Partner kooperieren im Hinblick auf deren spezifische Forschungsinteressen mit popgen in sog. Teilprojekten, innerhalb derer die Parameter für die Rekrutierung von Probanden definiert werden.

Der Betrieb von popgen als Biomaterialbank erstreckt sich von der Entgegennahme der Blutproben, die zusammen mit identifizierenden Daten der Probanden und i.d.R. weiteren medizinischen oder soziodemografischen Daten eingesandt werden, bis zur Analyse der aus den Blutproben hergestellten DNA-Proben und der Zusammenführung der genetischen mit den klinischen Daten zum Zwecke der wissenschaftlichen Auswertung im jeweiligen Teilprojekt.

2 Datenerhebung von den Probanden

Die Rekrutierung der sog. Patientengruppe erfolgt über eine Ansprache der erkrankten Personen im Einzugsgebiet des Projektes, vermittelt je nach Typ der Erkrankung über die örtlichen Behandlungszentren oder die behandelnden Ärzte. Die Probanden nehmen dann von sich aus Kontakt mit popgen auf; alle weiteren Schritte erfolgen auf der Grundlage einer Einwilligung, die von Teilprojekt zu Teilprojekt angepasst wird.

Daneben erfolgt auch die Rekrutierung von zufällig ausgewählten Personen auf der Grundlage von Daten der Meldeämter. Auf diese Weise kann die Häufigkeit bestimmter genetischer Varianten in einer repräsentativen Stichprobe im Projektgebiet festgestellt werden. Grundlage ist wiederum ausschließlich die Einwilligung der Betroffenen.

Während die identifizierenden Daten der Patientengruppe (gestützt auf die Einwilligungserklärung) für mind. 20 Jahre gespeichert bleiben, werden diese Daten bei der Kontrollkohorte nach abgeschlossenen Rekrutierung gelöscht.

3 Organisationsstruktur: Verteilung und Abschottung von Daten

Das Projekt popgen wird am Universitätsklinikum Schleswig-Holstein (UK-SH), Campus Kiel, von vier beteiligten Instituten durchgeführt.

Die Klinik für Allgemeine Innere Medizin betreibt die Kernressource von popgen, die sog. Studienzentrale. Dazu gehört auch das medizinisch-technische Eingangslabor, wo aus den eingehenden Blutproben DNA gewonnen wird.

Das Institut für Klinische Molekularbiologie (IKMB), eines der drei Genotypisierungszentren des NGFN, ist verantwortlich für die Lagerung, Verwaltung und Verarbeitung der DNA-Proben. Dabei wird ein Labor-Informations-und Managementsystem (LIMS) eingesetzt.

Das Rechenzentrum (RZ) des Universitätsklinikums betreibt den Pseudonymisierungsdienst (Trustee). Dieser ist ein Kernbestandteil des Konzepts der zweistufigen Pseudonymisierung, wie sie auch in [TMF06] vorgesehen wird.[1]

---

[1] [TMF06] S. 81.

Das Institut für medizinische Informatik und Statistik (IMIS) übernimmt die Zusammenführung von genetischen und klinischen Daten. Diese ist erforderlich für die wissenschaftliche Auswertung der Genotyp-Phänotyp-Beziehungen.

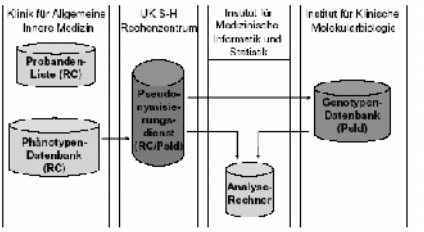

Abbildung 1: Beteiligte Organisationseinheiten

Nach der Systematisierung nach [TMF06] dürfte es sich bei popgen um eine eigenständige BMB mit zentraler Patientenliste handeln.[2] Um das Potential zur Rückidentifizierung zu minimieren, wird in [TMF06] empfohlen, mindestens die zentrale Datenbank (bei popgen: Phänotyp-Datenbank) und die Probenbank bei getrennten Personen zu führen, wobei die Patientenliste bei einer vertrauenswürdigen, unabhängigen Einrichtung liegen sollte.[3] An anderer Stelle[4] heißt es: „Die Patientenliste ist unbedingt räumlich und technisch getrennt von den Forschungsdaten angeordnet und auch einer getrennten disziplinarischen Verantwortung unterworfen. Es muss ein praktikables und tragfähiges Sicherheitskonzept vorliegen, das sicherstellt, dass die Unabhängigkeit gewährt ist.“ Diesen Anforderungen genügt popgen . Langfristig ist aber die weitere Perspektive im Auge zu behalten, wonach „bei besonders hohen Sicherheitsanforderungen (...) auch die Option (besteht), einen externen Datentreuhänder mit der Betreuung der Patientenliste zu beauftragen.“[5]

4 Probandenliste, Phänotypdatenbank, Recruitmentcode

Die zentrale administrative Datenbank für popgen befindet sich bei der Klinik für Allgemeine Innere Medizin, der sog. popgen Studienzentrale. Hier werden die identifizierenden Daten der Probanden in der Probandenliste gespeichert. Dies geschieht, wenn ein Teilnehmer sein Interesse mittels einer Rückantwortpostkarte äußert. Diesem Probanden wird dann das Testset zugesandt, bestehend aus dem Patientenaufklärungsbogen, der Einwilligungserklärung, einem Blutentnahmeset, dem Fragebogen um die relevanten soziodemografischen und ggf. medizinischen Daten zu erheben. Bereits zu diesem Zeitpunkt wird jedem Probanden in der Liste ein Pseudonym zugeordnet, der Recruitmentcode (RC).

---

[2] [TMF06] S. 71.
[3] a.a.O.
[4] [TMF06] S. 79.
[5] a.a.O.

Dieses maschinell erzeugte Pseudonym besteht aus zehn Zeichen, wobei die ersten drei Zeichen Buchstaben sind, die als Kürzel für die jeweilige Krankheitskohorte (gemäß den zuvor definierten Teilprojekten) bzw. die Zugehörigkeit zur Kontrollkohorte anzeigen. Die restlichen sieben Ziffern bestehen aus einer fortlaufend generierten Zahlenkombination. Der RC entspricht als erste Stufe der Pseudonymisierung der PID nach [TMF06].[6] Die Generierung von neuen RC ist durch technisch-organisatorische Maßnahmen auf die zuständigen Mitarbeiter beschränkt und wird protokolliert. Der RC wird auch als Barcode auf dem Fragebogen zu den medizinischen und soziodemografischen Angaben und auf den Blutprobenröhrchen aufgebracht, bevor diese an den Probanden versandt werden.

Nach Rücklauf des Testsets werden die medizinischen und soziodemografischen Daten sowie der Lagerort der Blutproben in die Phänotypdatenbank eingegeben. Als Zuordnung zu einem Probanden dient nur der RC, das erste Pseudonym. Sowohl Phänotypdatenbank als auch die Kühlschränke mit den Blutproben sind durch technisch-organisatorische Maßnahmen im Sinne von § 9 Bundesdatenschutzgesetz geschützt.

5 Weitere Pseudonymisierung: LabC, PsID

Nach der Extraktion der DNA-Proben aus den Blutproben wird die zweite Stufe der Pseudonymisierung vorgenommen. Dazu wird zunächst das Identifizierungskennzeichen der Proben umcodiert. Die DNA-Proben erhalten einen sog. Laborcode (LabC), der den RC ersetzt. Während der Präparation wird durch Zwischenspeicherung in einer Arbeitsdatei dafür gesorgt, dass ein vorläufiger Bezug zwischen RC und LabC hergestellt werden kann. Nach Abschluss der Extraktion wird diese Tabelle gelöscht und alle weiteren Spuren, die einen Bezug zum RC enthalten, vernichtet. Dies wird über eine Dienstanweisung abgesichert, die eine SOP festlegt.

Nach Abschluss der Präparation werden die Datensätze für die neu erzeugten DNA-Proben an den Pseudonymisierungsdienst geschickt. Der temporär in der popgen-Studienzentrale vorgehaltene Datensatz enthält neben den LabC und einer Ordnungsnummer zur Bezeichnung des Lagerorts der Proben noch den RC. Dieser wird im Pseudonymisierungsdienst in ein Pseudonym der zweiten Stufe, PsID, umgewandelt. Der pseudonyme Identifikator (PsID) ist ein zehnstelliger alphanumerischer String, bestehend aus sieben führenden Ziffern, die zufällig ausgewählt werden, und drei sich anschließenden Buchstaben, die die Kohortenzugehörigkeit angeben. Letztere werden im Rahmen der Pseudonymisierung dem Recruitmentcode entnommen und sind aus technischen Gründen notwendig.

[6] [TMF06] S. 79, 81.

Mit Hilfe einer im Pseudonymisierungsdienst verwendeten Datenbank, die alle bisher generierten und einem RC eindeutig zugewiesenen PsIDs speichert, ist gewährleistet, dass keine Zufallszahl als Bestandteil einer PsID doppelt vergeben wird und somit eine eindeutige Pseudonymisierung vollzogen werden kann. Gespeichert wird der PsID ausschließlich im LIMS des IKMB und in der Datenbank des Pseudonymisierungsdienstes.[7]

6 Pseudonymisierungsdienst

Der dedizierten Rechner im RZ des UK-SH, der die Pseudonymisierung vornimmt, ist über eine Firewall geschützt in das lokale Netz des RZ und damit auch in das Kliniknetz eingebunden ist. Inbox und Outbox des Pseudonymisierungsrechners liegen in der demilitarisierten Zone der Firewall. Sie sind von extern nur über Rechner mit spezifischen IPAdressen erreichbar. Zugriff von extern auf die Inbox hat nur der Rechner der popgen-Studienzentrale.

Auf dem Rechner im RZ läuft permanent ein Programm, das die Pseudonymisierung vornimmt. Dieses überprüft im Millisekundenabstand, ob neue Dateien in der Inbox eingetroffen sind, verarbeitet diese gegebenenfalls und löscht anschließend die Daten aus der Inbox. Dabei wird der RC jedes Datensatzes durch das zweite Pseudonym, den PsID, ersetzt. Wird ein RC hierbei erstmalig pseudonymisiert, so wird eine neue Zufallszahl generiert und überprüft, ob diese schon einmal vergeben wurde. Nach Zuweisung eines neuen PsID wird die Zuordnung dieses Pseudonyms zu einem RC in einer Datenbank auf dem Rechner gespeichert. Wurde für ein RC schon früher eine PsID vergeben, so wird die zugehörige PsID in der Datenbank gesucht. Jeder vorkommende RC wird so durch das ihm eindeutig zugewiesene Pseudonym ersetzt und der so pseudonymisierte Datensatz wird in ein anderes Verzeichnis, in die sog. Outbox geschrieben.

Die auf diesem Wege pseudonymisierte Datei kann nun von genau zwei über ihre IP-Adresse eindeutig spezifizierte Rechner über eine verschlüsselte Verbindung aus der Outbox in ein lokales Verzeichnis kopiert und anschließend auch dort gelöscht werden. Einen Rechner im IKMB, der den Import in das LIMS gewährleistet, und einen Rechner im IMIS, der im Rahmen der wissenschaftlichen Analyse für das Zusammenführen von Phänotypen und Genotypen genutzt wird.

Im IKMB werden die Exportdateien, welche vom Pseudonymisierungsdienst erzeugt wurden in das Labor-Informations- und Managementsystem (LIMS) importiert.

---

[7] Allein für die wissenschaftliche Analyse von Phänotyp und Genotyp verlassen die Pseudonyme das IKMB, dann allerdings nur temporär und an Orte, wo kein Rückschluss auf RC, persönliche oder medizinische Daten möglich ist.

7 Auswertung von Genotyp-Phänotyp-Beziehungen

Ziel des Forschungsprojekts popgen ist die wissenschaftliche Analyse von Genotyp-Phänotyp-Beziehungen. Dazu muss eine Liste von Genotypen derjenigen Individuen erstellt werden, die einen bestimmten Phänotyp haben, ohne dass Rückschlüsse auf persönliche oder medizinische Daten der Individuen möglich sind. Bei Vorliegen der prozeduralen Voraussetzungen für den Start eines neuen Teilprojektes werden die Daten derjenigen Probanden, auf die eine zuvor präzise definierte Phänotypkonstellation zutrifft, aus der Phänotyp-Datenbank bei der popgen-Studienzentrale selektiert. Solche Abfragen sind nur autorisierten Mitarbeitern möglich. Eine Dienstanweisung legt fest, dass als Ergebnis einer Abfrage immer mindestens 50 Datensätze ausgegeben werden müssen um das Risiko der Re-Identifizierung bei kleineren Gruppen zu minimieren. Ergebnis der Datenbankabfrage ist eine Liste von Phänotypen und den zugehörigen RCs, welche wiederum über den Pseudonymisierungsdienst weitergeleitet wird, wo die RCs in PsIDs umcodiert werden. In der Outbox des Pseudonymisierungsdienstes wird danach eine Liste von Phänotypen und zugehörigen pseudonymen Identifikatoren PsID abgelegt.

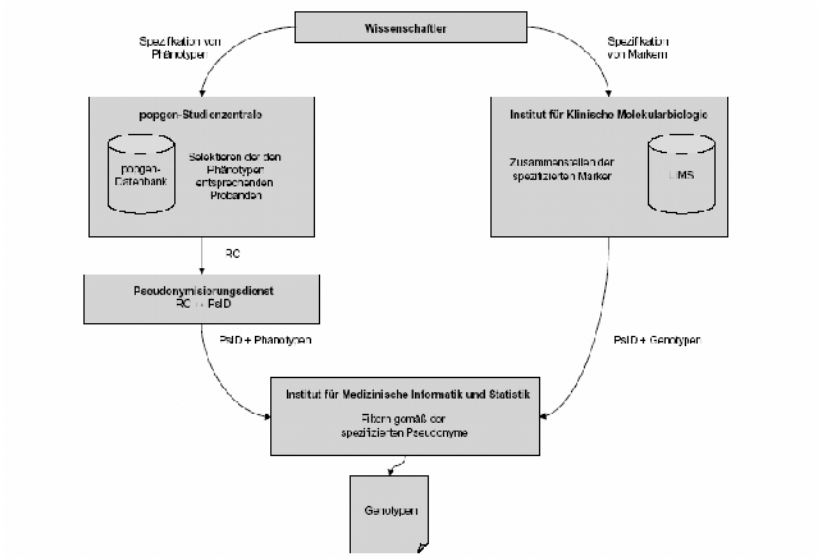

Abbildung 2: Zusammenführung pseudonymisierter Daten zur Auswertung

An dieser Stelle kommt die vierte organisatorische Einheit des UK-SH, das Institut für medizinische Informatik und Statistik (IMIS) ins Spiel. Über den in der Firewall erlaubten Zugriff eines dafür vorgesehenen Rechner im IMIS wird die Liste von der Outbox abgeholt und auf den Rechner im IMIS verschoben. Für die wissenschaftliche Analyse von Genotyp-Phänotyp-Beziehungen müssen jetzt die zugehörigen Genotypen den über den Phänotyp selektierten Datensätzen zugeordnet werden. Der Forscher beantragt beim IKMB eine Auflistung der Ausprägungen von spezifizierten Markern aller Individuen der interessierenden Kohorte. Das Ergebnis der Datenbankabfrage, eine über die PsIDs organisierte Auflistung von Genotypen, wird dem IMIS übermittelt.

Im IMIS liegt die Liste der pseudonymen Identifikatoren (PsIDs) jener Probanden vor, die die interessierenden Phänotypen aufweisen. Anhand dieser Liste werden aus der Gesamtheit der Kohorte die Datensätze von Genotypen ermittelt, die über den gesuchten Phänotyp verfügen. Nach dieser Operation werden aus der Liste der Resultate die PsIDs gelöscht, so dass eine faktisch vollständig anonymisierte Liste von Genotypen und Phänotypen entsteht. Allein diese Liste wird dem Forscher übergeben. Auf diese Weise gelangt weder eine Aufstellung von RCs noch von PsIDs in die Hände des Wissenschaftlers. Nach der Selektion werden alle während des Prozesses im IMIS anfallenden pseudonymen Daten permanent gelöscht. Damit wird der Anforderung in [TMF06] entsprochen[8], wonach der anonymisierte Datenexport an die Forscher immer dann zu wählen ist, wenn ein Rückbezug über die Pseudonyme für das Teilforschungsprojekt nicht nötig ist.

8 Kontrollgruppe

Den Schwerpunkt des Forschungsinteresses bilden Fall-Kontroll-Studien, in denen nach einer Assoziation eines Phänotyps mit gewissen genetischen Ausprägungen gesucht wird. Grundlegende Idee ist hier, dass unter der Annahme der Existenz einer solchen Assoziation gewisse Genotypen in der Gruppe derer, die den Phänotyp aufweisen, statistisch signifikant häufiger auftauchen als in der Kontrollkohorte.

Um den Forschern diese Kontrollproben zur Verfügung stellen zu können, werden auch Daten einer zufällig ausgewählten Gruppe von Bewohnern Nord-Schleswig-Holsteins erhoben. Anders als bei den Daten der Patientenkohorten kommt er hier aber nicht zu einer dauerhaften Speicherung von identifizierenden Daten. Namen und Anschriften dieser Kohorte wird nach Abschluss der Rekrutierung gelöscht; das Geburtsdatum wird in das Geburtsjahr verkürzt.

Da es nach dieser Anonymisierung keine Verbindung zwischen den soziodemografischen und medizinischen Daten eines Probanden der Kontrollgruppe und seinen identifizierenden Daten gibt, werden sowohl die Genotypen als auch die soziodemografischer und medizinischen Angaben (soweit sie erhoben werden), dauerhaft unter dem Recruitmentcode RC gespeichert. Eine weitere Pseudonymisierung ist hier nicht erforderlich.

Literaturverzeichnis

[TMF06] Projektgruppe Biomaterialbanken im TMF e.V.: Ein generisches Datenschutzkonzept für Biomaterialbanken, Version 1.0. April 2006

[8] S. 86 oben.