

# Datenschutz in der pharmakogenetischen Forschung – eine Fallstudie

Norbert Luttenberger, Joachim Reischl, Markus Schröder, Claus S. Stürzebecher

*Im Projekt GENOMatch wurde an der Universität Kiel in Zusammenarbeit mit der Schering AG ein Datenschutzkonzept für die pharmakogenetische Forschung entwickelt, das sowohl den Erfordernissen der Forschung als auch den Persönlichkeitsrechten der Probanden umfassend Rechnung trägt. Das Verfahren wurde vom ULD Schleswig-Holstein auditiert.*

Prof. Dr. Norbert  
Luttenberger

Professor am Institut  
für Informatik der  
Christian-Albrechts-  
Universität zu Kiel

E-Mail: nl@informatik.uni-kiel.de

Dr. Joachim Reischl

E-Mail: Joachim.Reischl@schering.de

Dipl.-Ing. Markus  
Schröder

E-Mail: schroeder@tembit.de

Dr. Claus Steffen  
Stürzebecher

E-Mail:

ClausSteffen.Stuerzebecher@schering.de

## Einleitung

Wann immer es um humangenetische Forschung geht, kommt dem Datenschutz eine besondere Bedeutung zu. Jedes Individuum hat eine unverwechselbare genetische Ausstattung; es gibt genetische Merkmale, die die Prognose bestimmter Erkrankungen erlauben, und viele genetische Merkmale sind nicht nur für das einzelne Individuum, sondern auch für seine Blutsverwandten aussagefähig. Genetische Merkmale lassen sich zu digitalen Daten aufbereiten, die von Rechensystemen gespeichert, verarbeitet und übertragen werden können. Jede Institution, die sich mit humangenetischen Daten beschäftigt, muss also für sich und ihre Klienten ein Konzept für den Schutz dieser Daten vor unberechtigtem Zugriff und missbräuchlicher Verwendung entwickeln und umsetzen.

In diesem Beitrag wird das Datenschutzkonzept vorgestellt, das im Rahmen eines Projekts mit dem Namen „GENOMatch“ kooperativ von der Schering AG und der Christian-Albrechts-Universität (CAU) zu Kiel für den Bereich der pharmakogenetischen Forschung der Schering AG entwickelt wurde. Mitbeteiligt an der Konzept- und der Softwareentwicklung war die TEMBIT Software GmbH in Berlin. Das Projektziel bestand darin, vor der eigentlichen Durchführung von pharmakogenetischen Studien einen „elektronischen Datentreuhänder“ zu verwirklichen, der durch besondere technische Merkmale und seine Einbettung in eine detailliert beschriebene organisatorische Struktur den Schutz der genetischen Daten von Teilnehmern an pharmakogenetischen Studien auf höchstmöglichem Niveau garantiert und gleichzeitig das Management der pharmakogenetischen Studien (inkl. des Managements der Blut- und Gewebeprobe) effizient unterstützt. Von Projektbeginn an war vorgesehen, das zu entwickelnde Konzept durch das Unabhängige Landeszentrum für Daten-

schutz Schleswig-Holstein (ULD) im Rahmen des sogenannten Behördenaudit – mit der CAU als konzeptvertretender „Behörde“ – auditieren zu lassen.

Da wir in diesem Aufsatz von Aktivitäten berichten, die vor der Aufnahme von realen pharmakogenetischen Studien durchgeführt wurden, werden keinerlei Aussagen zu den Inhalten einer von Teilnehmern zu unterzeichnenden Einwilligungserklärung gemacht („*Informed Consent*“). Selbstverständlich muss aber eine solche Einwilligungserklärung, die über den Verwendungszweck von Proben und Daten aufklärt, von den Studienteilnehmern eingeholt werden.

## 1 Datenschutz in der pharmakogenetischen Forschung

### 1.1 Was ist Pharmakogenetik?

Die Pharmakogenetik ist eine Spezialdisziplin der Genetik, deren Ziel es ist, den Einfluss genetischer Faktoren auf die differentielle Reaktion von Patienten auf Arzneimittel auf statistischer und individueller Ebene zu erforschen bzw. zu berücksichtigen. Sie findet Anwendung sowohl in der Entwicklung neuer Arzneimittel als auch (in einigen ausgewählten Gebieten) bei der Auswahl von bereits vorhandenen Therapieoptionen.

Bei der Arzneimittelforschung steht im Vordergrund, die unterschiedlichen Reaktionen von Patienten auf ein Arzneimittel mittels pharmakogenetischer Analysen besser oder überhaupt zu verstehen, sowohl im Hinblick auf Wirksamkeit und Wirkmechanismus als auch auf Nebenwirkungen. Die pharmakogenetischen Analysen, bei denen die klinischen Befunde mit genetischen Profilen abgeglichen werden, suchen dabei nach statistisch signifikanten Zusammenhängen zwischen genetischem Profil und Therapieantwort auf Ebene von Patienten.

tengruppen. Der Umfang der genetischen Analysen kann dabei von wenigen sogenannten Kandidatengenen bis zu ausführlichen genetischen Profilen reichen. Auf Basis der Ergebnisse solcher Analysen können idealerweise für Folgestudien Prädiktoren für den einzelnen Patienten ermittelt werden, z. B. im Hinblick auf den erwartbaren Behandlungserfolg, die als Auswahlkriterien für die Aufnahme in klinische Studien herangezogen werden. Im Fall schwerwiegender Nebenwirkungen könnten auch innerhalb einer laufenden Studie solche Rückschlüsse zum Schutz von Patienten relevant werden.

Mittelfristig können so die Erfolgsaussichten und die Kosten der klinischen Entwicklung neuer Arzneimittel positiv beeinflusst werden. Eine verbesserte Auswahl von Arzneimitteln nach pharmakogenetischen Parametern in der medizinischen Praxis kann mittel- bis langfristig eine individualisierte(re) Therapie ermöglichen.

Ein anderer wichtiger Aspekt der Pharmakogenetik in der klinischen Entwicklung ist die Möglichkeit, neue *targets* für Arzneimittel durch ein besseres Verständnis der Interaktion von Arzneimitteln und bestehender Krankheit zu identifizieren.

In der Pharmakogenetik sollen genetische Daten weder in der Forschungs-, noch in der Anwendungsphase eines Arzneimittels genutzt werden, um ein Individuum zu identifizieren. Die Pharmakogenetik verfolgt ganz andere Zielsetzungen als die Forensik, es geht nicht um die Erstellung eines *genetic fingerprint*. Im Forschungsprozess werden durch die biostatistische Auswertung der genetischen Daten einer großen Gruppe von Probanden/Patienten „unpersönliche“ Genprofile erstellt. Diese Genprofile sollen in der Anwendungsphase dazu dienen, die Therapieeignung einer Einzelperson zu prognostizieren. Dafür werden die von dieser Einzelperson erhobenen genetischen Daten mit dem statistisch ermittelten Genprofil abgeglichen.

## 1.2 Warum keine Anonymisierung genetischer Daten?

Metschke und Wellbrock weisen in [3] daraufhin, dass „Forschung mit sicher anonymisierten Daten [...] jederzeit ohne datenschutzrechtliche Vorgaben möglich [ist].“ In der pharmazeutischen Industrie werden unterschiedliche Datenschutzkonzepte angewendet, darunter auch solche, bei

denen (nach Vorgabe durch entsprechende interne Verfahrensanweisungen) der Bezug von Proben und Daten zu einem Patienten im o.g. Sinne irreversibel aufgehoben wird und damit kein Rückruf von Proben und auch kein Feedback bezüglich relevanter Resultate mehr möglich ist.

Wir haben bewusst ein Konzept favorisiert, das diese aus unserer Sicht wichtigen und wünschenswerten Optionen erlaubt und gleichzeitig den sich daraus ergebenden besonderen Anforderungen an den Datenschutz Rechnung trägt. Mitentscheidend war, dass wir den Empfehlungen einer Reihe von teils bereits geltenden, teils noch in Bearbeitung befindlichen Richtlinien zu genetischer Forschung (z. B. Europäische Datenschutzrichtlinie, UNESCO, CIOMS) und den sich abzeichnenden Forderungen von Datenschützern gerecht werden wollten, die möglicherweise in näherer Zukunft ihren Ausdruck in der Gesetzgebung finden werden (vgl. z. B. den Entwurf eines „Gesetzes zur Sicherung der Selbstbestimmung bei genetischen Untersuchungen“ der 62. Konferenz der Datenschutzbeauftragten des Bundes und der Länder, Oktober 2001 [1]).

Insgesamt sollte damit eine gewisse Zukunftsfähigkeit des Konzepts sichergestellt werden. Hinsichtlich der Abwägung zwischen Datensicherheit und Rückverfolgbarkeit von Proben gehen diese Empfehlungen und Forderungen in die Richtung, entweder die Anonymität zuzusichern oder aber das Recht auf Information und Rückzug der Proben sicherzustellen. Einige der Dokumente favorisieren die Anonymisierung (z. B. der oben zitierte Gesetzesentwurf [1]), andere das Recht auf Information und Erhalt des Bestimmungsrechts über die Proben und Daten aus genetischen Analysen (z. B. CIOMS)

Bevor einige weitere Gründe genannt werden, warum wir uns entschieden haben, Proben und Daten nicht zu anonymisieren, sondern „nur“ zu pseudonymisieren (was mit einem erheblich höheren Aufwand verbunden ist), sollen hier einige Vorbemerkungen zur Verwendung dieser Begriffe im Kontext der pharmakogenetischen Forschung gemacht werden.

Genetische Daten, die das Genom eines Individuums beschreiben, lassen sich wegen ihrer Einmaligkeit in letzter Konsequenz nicht in gleicher Weise anonymisieren wie andere körperliche Merkmale, z. B. Augenfarbe, Gewicht, Größe, Blutdruck usw. Präziser ausgedrückt: Für genetische Daten gibt es keine „Anonymitätsgruppe“ [4], in

der sich die genetische Ausstattung eines Individuums unter eine Menge von gleichen Ausstattungen „verstecken“ ließe, so dass ein externer Beobachter keine Chance hat, eine ganz bestimmte genetische Ausstattung einem ganz bestimmten Individuum zuzuordnen. In gleicher Weise wie der Fingerabdruck ist die genetische Ausstattung seinem Besitzer in einer 1-zu-1-Relation zugeordnet. (Wobei aber der Fingerabdruck keine „überschießende“ Information erschließt wie z. B. die einleitend erwähnten prognostischen Informationen.) Der Begriff Anonymisierung (und damit auch der Begriff Pseudonymisierung) ist also im Zusammenhang mit genetischen Daten in einem anderen Sinne zu gebrauchen.

Für unsere Zwecke folgen wir bei der Bestimmung des Begriffs Anonymisierung der *Recommendation No. R (97) 5 on the Protection of Medical Data of the Council Of Europe*: „The expression "personal data" covers any information relating to an identified or identifiable individual. An individual shall not be regarded as "identifiable" if identification requires an unreasonable amount of time and manpower. In cases where the individual is not identifiable, the data are referred to as anonymous.“ [2] Eine ähnliche Terminologie führt das *UNESCO International Bioethics Committee (IBC)* in seiner *Draft International Declaration On Human Genetic Data* (in der Fassung vom 8.10.2003) [6] ein. Es wird ausgeführt, dass genetische Daten an eine identifizierbare Person gebunden oder nicht gebunden sind (*linked / unlinked to an identifiable person*). Für nicht gebundene genetische Daten werden zwei Fälle unterschieden: Die Informationen zur Person des Probenspenders werden durch einen „code“ ersetzt (in unserer Sprechweise: ein Pseudonym), oder die Wiederherstellung der Bindung ist ausgeschlossen (*data irretrievably unlinked to an identifiable person*), da alle Informationen zur Person des Probenspenders gelöscht wurden. Im Art 14(d) des *Addendum 2* zu diesem Dokument (vom 8.10.2003) wird festgestellt: „Human genetic data, human proteomic data and biological samples collected for medical and scientific research purposes, can remain linked to an identifiable person, only if necessary to carry out the research and provided that the privacy of the individual and the confidentiality of the data or biological samples concerned is protected in accordance with domestic law.“

Im Sinne dieser beiden Ansätze sprechen wir von anonymisierten genetischen Daten,

wenn der Rückbezug dieser Daten zu einer identifizierbaren Person nur mit erheblichem Aufwand an Zeit und Arbeitskraft oder gar nicht möglich ist; für pseudonymisierte Daten gilt, dass die Herstellung des Rückbezugs im allgemeinen Fall den gleichen Aufwand bedeutet, dass allerdings in bestimmten Situationen und unter Beachtung genau beschriebener Prozeduren der Rückbezug „einfach“ hergestellt werden kann.

Diese Definitionen waren richtungweisend für die Entwicklung des hier beschriebenen Datenschutzkonzepts:

- Zum einen müssen genetische Proben und Daten so „behandelt“ werden, dass es tatsächlich nur mit erheblichem Aufwand möglich ist, den Rückbezug zu einer identifizierbaren Person herzustellen.
- Zum anderen müssen die Situationen und Prozeduren, die zu einer Herstellung des Rückbezugs führen, mit besonderer Umsicht beschrieben und durch entsprechende technische Maßnahmen abgesichert werden. Letztere müssen so gestaltet sein, dass sie aus heutiger Sicht ein sehr hohes Schutzniveau gewähren.

Auch wenn die Anonymisierung von genetischen Proben und Daten den datenschutzrechtlich bevorzugten Ansatz darstellt, wird – z. B. auch in [1] – anerkannt, dass wissenschaftliche Fragestellungen auch gegen eine Anonymisierung sprechen können. Darüber hinaus sind auch Aspekte des Patienteninteresses gegen die Vorteile der Anonymisierung abzuwägen. Aus den folgenden Überlegungen heraus ist deshalb eine Anonymisierung von genetischen Proben und Daten im Rahmen des GENOMatch-Projekts nicht vorgesehen:

- Wie bereits angeführt, wird Patienten das Recht eingeräumt, zu jeder Zeit zu verlangen, dass ihre Proben vernichtet und ihre Daten gelöscht werden sollen.
- Den Studienteilnehmern soll im Fall des ausdrücklich erklärten Wunsches eine Möglichkeit gegeben werden, Informationen über die eigene genetische Konstitution zu erhalten, sofern die erhobenen Daten validiert worden sind und von erheblicher Bedeutung für die Gesundheit sind.
- Sollen in einer Studie Arzneimittel erprobt werden, die bekannter Weise bei Patienten mit einem bestimmten genetischen Profil schwerwiegende Nebenwirkungen hervorrufen können, dann müssen Patienten auf Grund ihres genetischen Profils von der Teilnahme an sol-

chen klinischen Studien sicher ausgeschlossen werden können.

### 1.3 Integration des Datentreuhänders in das Study Management

Aus dem soweit Gesagten folgt, dass das Problem des Datenschutzes in der Pharmakogenetik (und ähnlichen Gebieten) anders gelöst werden muss als in einer Reihe von anderen Umfeldern, die datenschutzrelevante Probleme stellen (z. B. in der Telekommunikation, im e-Commerce usw.): In diesen Umfeldern können die betroffenen Personen ihre Identität schützen, in dem sie ihr persönliches *Identity Management* selber durchführen und damit die Aufdeckung ihrer Identität selber steuern.

In der Pharmakogenetik dagegen müssen dritte vertrauenswürdige Instanzen das *Identity Management* übernehmen. In der Pharmakogenetik beginnt das *Identity Management* in der klinischen Studie mit der Zuweisung einer Patienten-Nr. (PN) zu einer Person und schließt die „Fremd-Generierung“ von Identifikatoren für Blut- und Gewebeproben in der Biobank mit ein.

Eine Instanz, die das *Identity Management* für andere übernimmt, wird oftmals als „Datentreuhänder“ bezeichnet. Ein Datentreuhänder verfügt über die Identitätsdaten der beteiligten Personen, stellt sie anderen aber nur in wenigen, klar definierten Ausnahmefällen zur Verfügung. Im GENOMatch-Projekt war es das Ziel, einen elektronischen Datentreuhänder zu verwirklichen, der den durch die drei folgenden Qualitätsmerkmale charakterisierten Gesamtprozess effektiv unterstützt:

- Im eigentlichen pharmakogenetischen Forschungsprozess soll weder auf den Namen eines Patienten, der an einer pharmakogenetischen Studie teilnimmt, noch auf die bereits oben erwähnte Patienten-Nr. zurückgegriffen werden. Der Patientennamen ist allein dem Studienzentrum bekannt, die Patienten-Nr. wird bei der Einlagerung einer Blut-/Gewebeprobe von Probenträgern und Begleitinformationen entfernt.
- Der Prozess der Pseudonymisierung ist so zu gestalten, dass sich mindestens zwei Personen zusammenschließen müssen, um in missbräuchlicher Absicht eine einmal eingelagerte genetische Probe einer Patienten-Nr. zuzuordnen zu können. Wollen diese Personen gar den Namen des Probenpenders erfahren, müssen sie

sich mit einer entsprechenden Person im Studienzentrum „verbünden“.

- Um in missbräuchlicher Absicht von einem genetischen Datensatz auf eine Patienten-Nr. schließen zu können, müssen sich sogar mindestens drei Personen, die zwei verschiedenen Organisationen angehören, miteinander verbünden. Bezüglich des noch weitergehenden Rückschlusses auf den Patientennamen gilt das zuvor Gesagte.

Das Gesamtkonzept für einen solchen elektronischen Datentreuhänder muss technische und organisatorische Bedingungen festlegen, es sollte – am besten von unabhängiger sachverständiger Seite – kritisch überprüft werden, das zugehörige technische System ist mit aller Sorgfalt zu entwickeln und sollte zertifiziert werden, und vor allem sollten die für die Verwaltung der Identitätsdaten erforderlichen Datenbanken von unabhängiger dritter Seite betrieben werden.

Das GENOMatch-Datenschutzkonzept war unter der Randbedingung zu gestalten, dass es sich in ein Studienmanagementsystem sowohl konzeptuell als auch operativ nahtlos einfügen lassen würde. Eine wichtige Aufgabe des Studienmanagementsystems ist die Verwaltung von Blut- und Gewebeproben, die im Zuge von klinischen Studien eingelagert, aufbereitet, analysiert und nach Ablauf einer Frist oder bei Rücknahme der Patientenzustimmung vernichtet werden müssen. Eine weitere Aufgabe des Studienmanagementsystems ist die Verwaltung der aus den Proben erzeugten genetischen Daten. Anders ausgedrückt: Das Studienmanagementsystem ist so geplant worden, dass der Datenschutz von vornherein mitbedacht wurde. Dadurch entsteht zwar für den Datenschutz nach wie vor ein Mehraufwand; dieser ist jedoch gering, gemessen an dem Aufwand, der zu erwarten wäre, wenn Datenschutzmaßnahmen zu einem späteren Zeitpunkt zu einem existierenden Studienmanagementsystem hinzuzufügen wären.

## 2 Das GENOMatch-Projekt

Das globale Ziel des GENOMatch-Projekts ist die Herstellung einer IT-Infrastruktur für die pharmakogenetische Forschung der Schering AG, wobei aus den oben dargestellten Gründen dem Datenschutzkonzept eine besondere Bedeutung zukommt.

## 2.1 GENOMatch-Phasenmodell

Um die Komplexität des GENOMatch-Projekts beherrschbar zu machen, wurden für GENOMatch drei Arbeitsphasen definiert:

### Phase 1 – *Sample and save*:

Gegenstand dieser Phase ist die Definition derjenigen organisatorischen Abläufe und zugehörigen technischen Komponenten des Studienmanagementsystems, die die Aufbewahrung und das Handling von Blut- und Gewebeproben betreffen. Dies schließt ein die Definition von Verfahren für

- ◆ die sichere und pseudonyme Einlagerung von Proben in eine Biobank,
- ◆ den Versand von Proben aus der Biobank an bestimmte Labors,
- ◆ die Zuordnung von genetischen Daten zu Proben, und
- ◆ den Ausstieg eines Teilnehmers aus einer Studie (inkl. der Vernichtung von Proben und der Löschung von genetischen und probenbezogenen Daten).

Obwohl in dieser Phase noch keine genetischen Daten erzeugt werden, ist sie unter Datenschutzaspekten sicherlich die wichtigste, da hier die Grundlage für alle weiteren Verfahrensschritte gelegt wird.

### Phase 2 – *Secure storage of pseudonymized genetic and clinical data*

In dieser Phase wird die angesprochene IT-Infrastruktur erweitert um Komponenten für die Speicherung genetischer Daten und die Aufbereitung klinischer Daten, die ja in den pharmakogenetischen Forschungsprozess ebenso wie die genetischen Daten eingehen. Dabei wird vor allem dafür gesorgt, dass eine Re-Identifizierung der genetischen Daten über bestimmte Komponenten der klinischen Daten, die oftmals „sprechend“ sind, nicht erfolgen kann.

### Phase 3 – *Biostatistical Processing and aggregated data reporting*

In dieser Phase wird die aufgebaute Infrastruktur für verschiedene pharmakogenetische Analysen genutzt. Unter Datenschutzaspekten ist von besonderem Interesse, wie Ergebnisse von pharmakogenetischen Studien berichtet werden, d. h. z. B. wie Daten aggregiert werden.

## 2.2 Beteiligte Institutionen

An der pharmakogenetischen Forschung sind mehrere Arten von unabhängigen

Organisationen beteiligt, die ggf. in mehreren Instanzen vorhanden sein können:

- der Auftraggeber einer pharmakogenetischen Studie, der neben Aufsichtsaufgaben auch die Auswertung von klinischen und genetischen Daten in einem eigenen Rechenzentrum übernimmt (in unserem Fall ist diese Rolle durch die Fa. Schering AG besetzt),
- das Studienzentrum, d. h. die Klinik, in der zu einer laufenden klinischen Arzneimittelstudie eine „parallele“ pharmakogenetische Studie aufgesetzt wird, und von der Blut- und Gewebeproben einerseits und klinische Daten andererseits nach entsprechender Patienteneinwilligung für die pharmakogenetische Forschung zur Verfügung gestellt werden,
- die Biobank, die für die Lagerung und Verwaltung von Blut- und Gewebeproben zuständig ist,
- das DNA-Extraktionslabor, das DNA aus den Blut- und Gewebeproben extrahiert,
- das DNA-Analyselabor, das die gewonnene DNA sequenziert und genetische Daten zur Verfügung stellt,
- das Rechenzentrum für die biostatistische Auswertung von klinischen und genetischen Daten. (Dieses Rechenzentrum wird – wie oben ausgeführt – vom Auftraggeber selber betrieben, oder es kann sich um ein unabhängiges externes Rechenzentrum handeln, wenn dieses über die notwendigen Einrichtungen für den Schutz der genetischen Daten verfügt.)

Diese Aufzählung macht deutlich, dass das Datenschutzproblem in der pharmakogenetischen Forschung allein schon deshalb schwierig zu lösen ist, weil mehrere unabhängige Organisationen eingebunden werden müssen, die sich an die „Spielregeln“ zu halten haben. U.a. deshalb ist es wichtig, die o. g. Rolle des Datentreuhänders möglichst effizient zu realisieren.

Im Folgenden wird das Hauptaugenmerk auf die erste GENOMatch-Arbeitsphase *sample and save* (siehe 2.1) gerichtet. Das Konzept für diese Projektphase ist detailliert ausgearbeitet und wurde vom ULD auditiert.

## 3 Technische Maßnahmen

### 3.1 Pseudonyme, Proben- und Datensatz-Identifikatoren

Innerhalb einer klinischen Studie werden Patienten über eine sog. Patienten-Nr. (PN) referenziert. Jeder Studie wird im erwähnten Studienmanagementsystem eine sog. Studien-Nr. (SN) zugeordnet. Damit bildet das Tupel (SN, PN) ein eindeutiges Patienten-Pseudonym. Da dieses Pseudonym de facto oftmals leicht zum Patientennamen hin aufzulösen ist, gilt es in der pharmakogenetischen Forschung als schützenswert. Es wird deshalb bei der Einlagerung von Proben ersetzt durch eine sog. *Sample Group Number* (SGN), die als Primärschlüssel zum Zugriff auf alle Probenidentifikatoren dient. Die SGN ist eine 12-stellige dezimale Zufallszahl. Wie der Name erkennen lässt, wird die SGN weniger als Personenpseudonym verstanden denn als Schlüssel für den Zugriff auf eine „Gruppe“ von genetischen Proben, die von einer Person stammen.

Von jedem Teilnehmer an einer pharmakogenetischen Studie gehen eine oder mehrere Blut- und/oder Gewebeproben und ein Datensatz mit klinischen Daten in den Forschungsprozess ein. Selbstverständlich muss es für die biostatistische Auswertung möglich sein, die Blut-/Gewebeproben bzw. die daraus gewonnenen genetischen Daten den klinischen Daten eindeutig zuzuordnen.

Die von den Studienzentren an die jeweilige Biobank gesandten Blut-/Gewebeproben sind mit dem bereits oben erwähnten (SN, PN)-Pseudonym bezeichnet. Diese Verfahrensweise erlaubt es einerseits, im Studienzentrum die aus der klinischen Forschung bekannten Prozesse weitestgehend beizubehalten, und andererseits kann die Biobank bestimmte Plausibilitätskontrollen vor der eigentlichen Einlagerung der Proben durchführen und ggf. auftretende Unstimmigkeiten in Zusammenarbeit mit dem liefernden Studienzentrum ausräumen. Nach Durchführung dieser Kontrollen werden von der Biobank die personenbezogenen Identifikatoren gegen Probenidentifikatoren ausgetauscht, die aus einer 12-stelligen dezimalen Zufallszahl und einer 2-stelligen Prüfsumme gebildet werden. Dieser Austausch wird in zwei Schritten vollzogen.

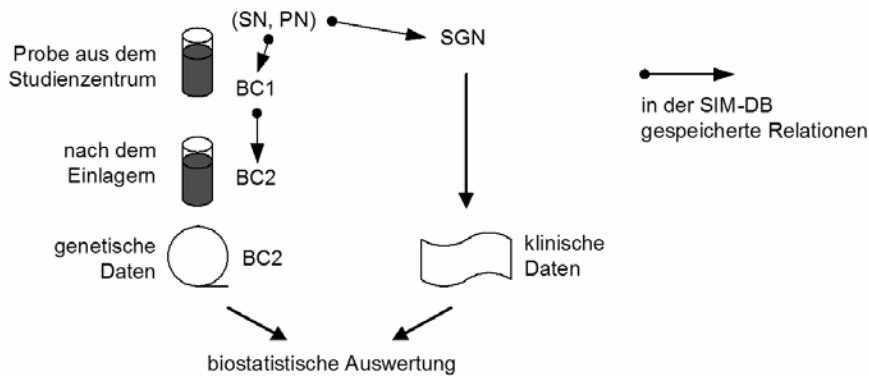


Bild 1: Weg von Proben und Daten

Da die Probenidentifikatoren als Strichcodes auf den Probenröhrchen angebracht werden, werden sie hier der Anschaulichkeit halber als *Barcodes* (BC) bezeichnet. Die Stellenanzahl für die Barcodes ist im Wesentlichen dadurch begrenzt, dass ein solcher Barcode auf ein sehr kleines „Fähnchen“ gedruckt werden muss, mit dem das Probenröhrchen etikettiert wird.

Beim Austausch registriert der jeweilige Biobankmitarbeiter die Zuordnung der Identifikatoren zueinander bei einer so genannten *Secure Identifier Management Database* (SIM-DB). Diese SIM-DB wird von einem unabhängigen Provider betrieben und ist mit wirkungsvollen Sicherheitsvorkehrungen vor unberechtigtem Zugriff geschützt. Gemeinsam mit der zugehörigen Datenbankapplikation bildet sie die Kernkomponente des elektronischen Datentreuhänders.

### 3.2 Entfernen identifizierender Daten

Bevor klinische Daten in den pharmakogenetischen Forschungsprozess eingespeist werden, werden alle Datensatzkomponenten entfernt, die die Identifikation des Studienteilnehmers erleichtern würden. Insbesondere werden die Teilnehmer-Initialen aus den klinischen Datensätzen gelöscht, und die Geburtsstagsdaten werden durch Altersangaben ersetzt.

### 3.3 Datenaggregation

Die biostatistische Auswertung erzeugt aggregierte Ergebnisse, die keinen Rückschluss auf einen einzelnen Studienteilnehmer erlauben.

## 4 Organisatorische Prozesse

Das gesamte Datenschutzkonzept basiert auf einem Zusammenspiel zwischen Handlungsanweisungen (*Standard Operating Procedures*, SOPs) und technischen Lösungen. Dabei wurden überall dort, wo technische Lösungen zur Erhöhung der Sicherheit implementierbar sind, auf diese zurückgegriffen und möglichst wenig ausschließlich durch Handlungsanweisungen/persönliche Verantwortlichkeiten geregelt.

Um die reibungslose Umsetzung des Konzepts zu ermöglichen, wurden bereits in einem sehr frühen Stadium der Entwicklung die späteren Nutzer des GENOMatch-Systems mit einbezogen. GENOMatch basiert auf einem elaborierten Rollenkonzept, das im Folgenden erläutert wird.

### 4.1 Studienzentrum

Im Studienzentrum ist der Studienarzt verantwortlich für die Einholung eines *Informed Consent* von Patienten, die Willens sind, an einer pharmakogenetischen Studie teilzunehmen, für die korrekte Überstellung von Proben an die Biobank und für die Meldung solcher Studienteilnehmer, die beschließen, aus einer Studie auszusteigen. Den letztgenannten Fall beschreiben wir in Pkt. 4.5.

Der Studienarzt versieht jeden Probenbehälter, in den er eine Blut- oder Gewebeprobe eines Studienteilnehmers einfüllt, mit zwei Etiketten: Das eine Etikett trägt das den Patienten identifizierende (SN, PN)-Tupel, das andere einen BC1. Das BC1-Etikett wählt der Studienarzt willkürlich aus einer Menge von BC1-Etiketten aus, die vor Beginn einer Studie an das Studienzentrum geliefert worden sind. Weiterhin erstellt der Studienarzt für jeden Studienteilnehmer

einen sog. *Pharmacogenetic Accompanying Letter* (PAL). In diesem bestätigt er gegenüber der Biobank, dass der Studienteilnehmer einen *Informed Consent* unterzeichnet hat. Das Original dieses PAL, in dem der Studienteilnehmer per Name identifiziert wird, verbleibt im Studienzentrum. Auf dem PAL-Original bringt der Studienarzt außerdem Dubletten der für den jeweiligen Studienteilnehmer verwendeten BC1-Etiketten an. Eine Kopie des PAL, in dem der Studienteilnehmer nur über das (SN, PN)-Tupel (und indirekt über eine weitere Plausibilitätsinformation, z. B. das Geburtsdatum) identifiziert wird (und auf das keine BC1-Dubletten aufgeklebt sind), gelangt mit den Proben des Studienteilnehmers an die Biobank. Da die Proben ebenfalls per (SN, PN) identifiziert sind, kann sich die Biobank (genauer gesagt: der sog. *Sample Registrar*, s.u.) davon überzeugen, dass für die jeweilige Probe ein *Informed Consent* vorliegt.

### 4.2 Biobank

Die für den Biobank-Betreiber definierten organisatorischen Prozesse schließen vier Rollen ein, die von verschiedenen Personen wahrgenommen werden müssen:

#### ■ *Sample Registrar*

Der *Sample Registrar* unterzieht die Proben, die vom Studienzentrum geliefert werden, zunächst einer Eingangsprüfung (Unversehrtheit der Probe, Plausibilität der Begleitdaten, Vorliegen eines korrekten PAL), registriert den auf der Probe angebrachten BC1 bei der *Secure Identifier Management Database*, entfernt den personenbezogenen Identifikator (SN, PN) von der Probe und legt die Probe in einem Gefrierschrank ab, der als Zwischenstation dient. Die PAL-Kopie wird vom *Sample Registrar* im Archiv der Biobank verwahrt und ist nur ihm zugänglich.

#### ■ *Sample Code Exchanger*

Der *Sample Code Exchanger* hat die einzige Aufgabe, den BC1 gegen einen BC2 auszutauschen. Der BC2 wird ebenfalls bei der *Secure Identifier Management Database* registriert. Der *Sample Code Exchanger* lagert die Probe dann in einem Gefrierschrank ein.

Durch diesen zweiten Tausch des Probenidentifikators kann das oben angegebene Ziel erreicht werden: Mindestens zwei Personen müssen sich zusammenschließen, um in missbräuchlicher Absicht eine genetische Probe einer Patienten-Nr. zuordnen zu können. Und zwar müssten – gegen die

vorgesehenen Regeln – sowohl der *Sample Registrar* als auch *Sample Code Exchanger* alle Austauschvorgänge in zwei „Nebenprotokollen“ erfassen, so dass sie zusammen aus diesen Protokollen die Zusammenhänge (SN, PN) ↔ BC1 ↔ BC2 rekonstruieren könnten. Dies wird allein schon dadurch erschwert, dass die BC2-Etiketten keine lesbare Nummer, sondern nur den Strichcode tragen.

#### ■ *Sample Manager*

Vom oben erwähnten zweiten Kühlschranks greift der *Sample Manager* auf Blut- und Gewebeproben zu. Er hat die Aufgabe, nach Anforderung durch den Auftraggeber der Studie dem DNA-Extraktionslabor Proben zuzusenden, die extraktierte DNA einzulagern, DNA an das DNA-Analyselabor zu senden und wieder einzulagern und eine Buchführung über die jeweils vorhandenen Bestände zu führen. Die Probenidentifikatoren werden vom *Sample Manager* nicht verändert. DNA-Proben tragen den gleichen BC2-Identifikator wie die Blut-/Gewebe-proben, aus denen sie erzeugt worden sind. Auch die genetischen Datensätze, die vom DNA-Analyselabor erzeugt werden, tragen diesen Identifikator.

#### ■ Administrator

Der Administrator weist verschiedenen Personen die genannten Rollen zu.

### 4.3 DNA Extraction und DNA Analysis Laboratory

Die von diesen beiden Laboratorien zu erledigenden Aufgaben haben keine Implikationen für den Datenschutz. Sie „sehen“ beide nur Proben bzw. Daten, die mit einem BC2 etikettiert sind. Der Klarheit halber sei darauf hingewiesen, dass das *DNA Analysis Laboratory* die gewonnenen genetischen Daten nicht an die Biobank sendet, sondern an den Auftraggeber der Studie, genauer gesagt: in die *Secure Data Area* (s. u.) des Auftraggebers, der sie dort langfristig speichert.

### 4.4 Pharmacogenetic Data Analysis Facility

Die *Pharmacogenetic Data Analysis Facility* wird vom Auftraggeber der Studie und/oder von einer unabhängigen Organisation betrieben. Sie hat die Aufgabe, klinische und genetische Daten gemeinsam mit Hilfe von biostatistischen Verfahren auszuwerten. Bei der Definition der organisatorischen Abläufe für die *Pharmacogenetic*

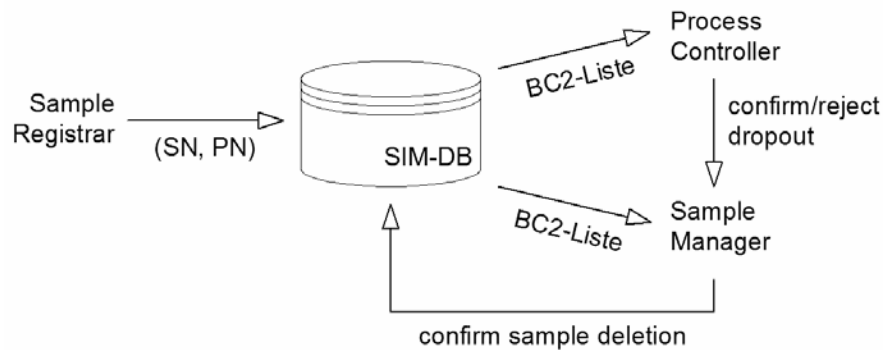


Bild 2: Biobank-Aktionen nach Rückzug eines Informed Consent

*Data Analysis Facility* war das Problem zu lösen, die mit einem (SN, PN)-Tupel identifizierten klinischen Daten eines Teilnehmers mit seinen genetischen Daten zusammenzubringen, ohne das (SN, PN)-Tupel bekanntzugeben, das ja wie oben ausgeführt in unserem Ansatz als schützenswertes Merkmal gilt.

Vor der biostatistischen Auswertung werden die klinischen Datensätze in zwei Schritten aufbereitet, die beide nur von der *Pharmacogenetic Data Analysis Facility* des Auftraggebers durchgeführt werden können: Im ersten Schritt werden – wie in 4.2 beschrieben – identifizierende Komponenten aus diesen Datensätzen entfernt, und dann wird das (SN, PN)-Tupel durch eine SGN ersetzt (vgl. 4.1). In dieser Form gelangen die klinischen Daten in eine für die biostatistische Auswertung einzurichtende sog. *Secure Data Area* und werden dort für die Dauer der Datenauswertung gespeichert. Von der *Secure Data Area* aus kann über die SGN auf die BC2-Proben-Identifikatoren und damit auf die genetischen Daten zugegriffen werden. Die Verbindung zwischen SGN und BC2-Identifikatoren wird über eine Anfrage bei der *Secure Identifier Management Database* hergestellt.

Von den für die *Pharmacogenetic Data Analysis Facility* definierten Rollen sind unter Datenschutz-Aspekten nur zwei interessant: der sog. *Process Controller* und der Administrator. Der *Process Controller* richtet im Studienmanagementsystem klinische Studien ein: Vergabe von Studiennr. (SN) und Studienname, zugehörige Studienzentren und Biobank(en). Außerdem hat er den gesamten mit der Durchführung von pharmakogenetischen Studien verbundenen Arbeitsablauf auf korrekte Durchführung zu überwachen. Dies schließt auch gewisse Schritte beim Ausstieg von Teilnehmern aus einer Studie ein, s. u. Die

Rolle des Administrators ist ähnlich definiert wie bei der Biobank: Er ist für die Zuweisung von Personen zu Rollen verantwortlich.

### 4.5 Ausstieg eines Teilnehmers aus einer Studie

Besondere Regeln gelten, wenn ein Teilnehmer aus einer pharmakogenetischen Studie aussteigen will. In diesem Fall müssen alle zugehörigen Proben vernichtet und Daten gelöscht werden (offensichtlich mit Ausnahme solcher Daten, die in aggregierter Form bereits in eine abschließende Analyse eingeflossen sind). Ein Studienteilnehmer, der aus einer Studie aussteigen will, wendet sich an seinen Studienarzt. Dieser übermittelt der Biobank in einem förmlichen Schreiben das (SN, PN)-Tupel des ausstiegswilligen Teilnehmers.

Der *Sample Registrar* teilt der SIM-DB dieses (SN, PN)-Tupel mit. Daraufhin versendet die SIM-DB Listen mit den zugehörigen BC2s sowohl an den *Process Controller* als auch an den *Sample Manager*. Der *Sample Manager* hat die Aufgabe, nach Bestätigung des Ausstiegs durch den *Process Controller* die identifizierten Proben zu vernichten und der SIM-DB die Vernichtung zu bestätigen. Zur Löschung evtl. bereits vorhandener genetischer Daten wird in ähnlicher Weise vorgegangen.

## 5 Datenschutzmanagementsystem

Unter dem Begriff „Datenschutzmanagementsystem“ wird üblicherweise die Gesamtheit aller organisatorischen und rechtlichen Regelungen begriffen, die notwendig sind, um das in einem Datenschutzkonzept beschriebene Gesamtsystem über die gesamte Betriebsperiode hinweg einsatzfähig

zu erhalten, in regelmäßigen zeitlichen Intervallen zu überprüfen und darin ggf. auftretende Fehlfunktionen zu erkennen und abzustellen. Es schließt die eindeutige Definition von Verantwortlichkeiten, die Schulung der beteiligten Personen, die Festlegung von Maßnahmen bei Unregelmäßigkeiten, die Erzeugung von *Audit Trails* usw. ein. Bezüglich eines solchen Datenschutzmanagementsystems wurden im GENOMatch-Projekt die üblichen Regelungen getroffen.

## 6 Implementierung

Für das GENOMatch-Projekt bestand eine wichtige Konzeptentscheidung darin, das System zum Management von Identifikatoren (PN sowie probenbezogene Identifikatoren) von Anfang an als Subsystem des Studienmanagementsystems aufzufassen – ein Subsystem, das zwar mit besonderen Vorkehrungen für sichere Authentifizierung und Autorisierung versehen ist, aber dennoch eng mit dem Studienmanagement verzahnt ist. Anders ausgedrückt: Das Studienmanagementsystem benötigt aus offensichtlichen Gründen einen einfachen und effizienten Zugriff auf die verwendeten Identifikatoren. Wenn man die entsprechenden Datenbestände durch sichere Zugriffsfunktionen (d. h. einen „elektronischen Datentreuhänder“ wie oben erwähnt) kapselt, ist als Resultat ein Studienmanagementsystem zu erwarten, das sowohl leistungsfähig und wirtschaftlich ist als auch besondere Qualitäten in puncto Datenschutz aufweist. Wir erläutern im Folgenden das Konzept.

### 6.1 Systemkonfiguration

Die Hardware des Studienmanagementsystems besteht aus zwei leistungsfähigen PCs; der eine dient als *Application Server*, der andere als *Database Server*. (Der Einsatz weiterer Maschinen zur Spiegelung oder Verteilung der verwendeten Datenbanken wurde in Erwägung gezogen, aus Kosten-Nutzen-Gründen wurde jedoch die einfachere Lösung realisiert.) Die beiden Server sind miteinander über ein eigenes LAN verbunden. Der *Application Server* ist zusätzlich über ein Firewall-System mit dem Internet verbunden, und der *Database Server* ist an ein Backup-System angeschlossen.

Das Gesamtsystem wird von *Dataport* betrieben, einer zertifizierten Anstalt des

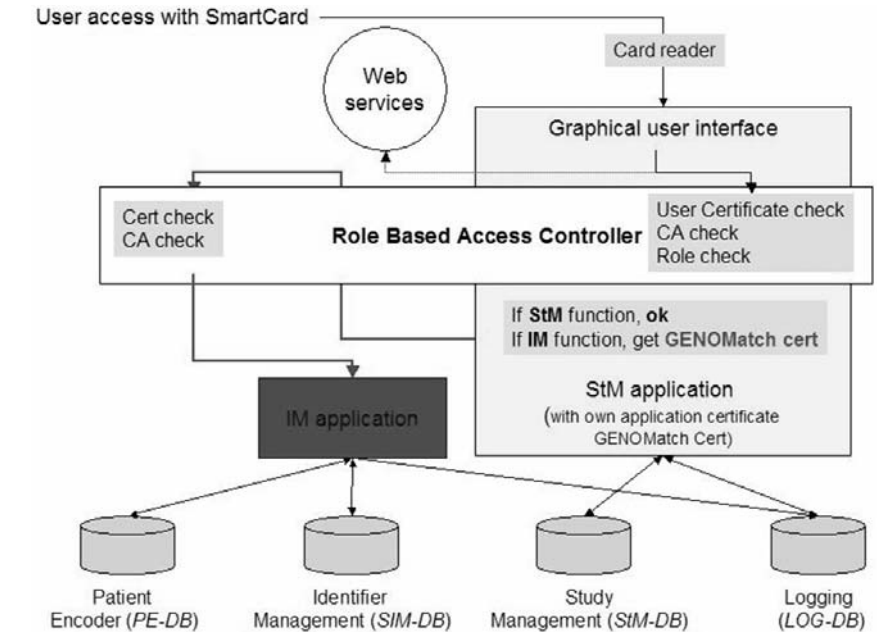


Bild 3: GENOMatch-Softwaresystem

öffentlichen Rechts, deren Träger das Land Schleswig-Holstein sowie die Freie und Hansestadt Hamburg sind. Der Sitz von *Dataport* ist Kiel-Altenholz. Die Maschinen werden von *Dataport* in einem streng zugangskontrollierten unterirdischen Raum betrieben, in dem für konstante Klimabedingungen und eine unterbrechungsfreie Stromversorgung gesorgt ist. *Dataport* ist für die Durchführung eines regelmäßigen Backup verantwortlich und bietet auch die Dienste einer *Certification Authority* (CA) an, die im Rahmen des GENOMatch-Projekts genutzt werden.

Als Client-Systeme dienen HTML-Browser. In der Biobank laufen die HTML-Browser auf sog. *Thin Clients*. *Thin Clients* sind Rechensysteme, die nur für eine genau definierte Applikationssoftware genutzt werden können; die Benutzer selber können keine weitere Software installieren, auch nicht über das Netz. Der Hintergrund für die Benutzung von *Thin Clients* in der Biobank ist folgender: In der Biobank werden – wie oben dargestellt – personen- und probenbezogene Identifikatoren eingelesen und registriert. Könnte ein missbrauchswilliger Biobank-Mitarbeiter auf den Eingabestationen ein „Sniffer“-Programm installieren, dann könnte er sich darüber in den Besitz aller Identifikatoren und ihrer Zusammenhänge bringen und diese rechtswidrig nutzen. Damit wäre die Funktion des elektronischen Datentreuhänders ausgehebelt.

Benutzer müssen sich gegenüber dem Studienmanagementsystem mit einer

SmartCard authentifizieren. Dazu sind die entsprechenden Geräte mit SmartCard-Lesern ausgestattet.

### 6.2 Softwaresystem

Das Applikationssystem besteht aus drei Hauptkomponenten (Bild 3):

- ♦ der *Studienmanagement*-Applikation (StM-Applikation)

Die StM-Applikation stellt Funktionen für die Verwaltung von GENOMatch-Benutzern, Studien (z. B. zugehörige Studienzentren, Biobanken, Studienstatus) und für die Verfolgung von Proben (z. B. Probenstatus, Füllstände usw.) zur Verfügung.

- ♦ der *Identifier Management*-Applikation (IM-Applikation)

Die IM-Applikation ist die Datenbank-Applikation zu der bereits erwähnten SIM-DB. Die IM-Applikation verwaltet alle personen- und probenbezogenen Identifikatoren und ihre Zusammenhänge. Über die IM-Applikation kann z. B. der *Sample Registrar* die BC1-Codes registrieren, die zu einem (SN, PN)-Tupel gehören. Die technische Realisierung der IM-Applikation schließt aus, dass menschliche Benutzer direkt auf die IM-Applikation zugreifen können; stattdessen erfolgt der Zugriff indirekt über die StM-Applikation. Diese muss sich gegenüber der IM-Applikation mit einem eigenen GENOMatch-Applikationszertifikat (GA) authentifizieren, das wiederum den Namen des Benutzers als Parameter mit sich führt. (Details

zur Zugriffskontrolle s. u.) Durch diesen „Umweg“ wird der direkte Durchgriff vom Browser auf die IM-Applikation ausgeschlossen. Zusätzlich stellt die IM-Applikation der Biobank und dem Rechenzentrum für die biostatistische Auswertung einige ausgewählte Dienste zur Verfügung. Diese Dienste werden in Form von *Web Services* über eine https-Verbindung angeboten.

◆ der Applikation für die rollenbasierte Zugriffskontrolle (*Role-based Access (RBA) Controller*)

Der RBA Controller überwacht den Zugriff der GENOMatch-Benutzer zu den Funktionen der StM- und der IM-Applikation. Dazu gelten die folgenden Voraussetzungen: Die GENOMatch-Benutzer nehmen Rollen im Sinne des dargestellten Rollenkonzepts wahr. Jedem Benutzer ist genau eine Rolle zugewiesen. (Nur der lokale Administrator kann eine Doppelfunktion einnehmen.) Außerdem verfügt jeder Benutzer über eine SmartCard mit einem persönlichen X.509-Zertifikat. Die GENOMatch-Funktionen werden über unterschiedliche URLs identifiziert. GENOMatch-Benutzer greifen auf diese Funktionen über https zu, wobei der zugehörige SSL-Tunnel beim RBA Controller endet. Nach Prüfung der Gültigkeit des Zertifikats und der Ermittlung des Benutzernamens (sog. *Distinguished Name* im Zertifikat) überprüft der RBA Controller, ob die Rolle, die der Benutzer einnimmt, die aufgerufene Funktion aufrufen darf. Bei einer positiven Antwort werden Anfragen an die StM-Applikation direkt von dieser Applikation ausgeführt, während Funktionsaufrufe an die IM-Applikation zusammen mit dem GENOMatch-Applikationszertifikat noch einmal zum RBA-Controller geschickt werden, wo sie gemeinsam geprüft und dann erst dann an der IM-Applikation zur Ausführung weitergeleitet werden. Für die IM-Applikation ist somit eine zweistufige Zugriffskontrolle implementiert, so dass Angriffe nicht autorisierter Benutzer auf die SIM-DB mit großer Sicherheit zum Scheitern verurteilt sind.

Auf dem Database Server sind vier Datenbank-Schemata installiert:

◆ die *Patient Encoder Database*

Diese Datenbank enthält die Patientendaten (PN, SN, Plausibilitätsinformation, *Sample Group Number* (SGN) und *Dropoutflag*). Sie ist intern mit einem symmetrischen Verschlüsselungsverfahren verschlüsselt.

◆ die *Secure Identifier Management Database*

In dieser Datenbank sind alle probenbezogenen Identifikatoren und ihre Zusammenhänge gespeichert. Die SGN fungiert als Primärschlüssel. Auch diese Speicherung (SGN, BC1, BC2) erfolgt in verschlüsselter Form. Nur die IM-Applikation kann auf diese Tabelle zugreifen.

■ die *Study Management Database*

Diese Datenbank enthält Tabellen mit Angaben zu den GENOMatch-Benutzern und ihren Rollen, mit allen Studieninformationen, mit Informationen zu Biobanken und mit Daten für die Verfolgung von Proben. Ausschließlich die StM-Applikation kann auf diese Datenbank zugreifen.

■ die *Logging Database*

In dieser Datenbank werden alle Zugriffe von GENOMatch Benutzern mitprotokolliert. Die erfassten Daten dienen als Datenquelle für den internen Audit und den *Authority Audit*.

## Fazit

Der Kern des dargestellten technisch/organisatorischen Datenschutzkonzepts besteht in einem mehrstufigen Pseudonymisierungsverfahren. Dieses Konzept wurde dem Unabhängigen Landeszentrum für Datenschutz Schleswig-Holstein unter Federführung durch die Christian-Albrechts-Universität zu Kiel zur Auditierung vorgelegt. Das ULD-Datenschutzaudit bestätigt, dass

dieses „in sich schlüssige und umfassende Konzept ... geeignet [ist], das Vertrauen von Probanden in den langfristigen Schutz der eigenen personenbezogenen Daten im Rahmen der pharmakogenetischen Arzneimittelforschung zu stärken“.

Eine wichtige Herausforderung wird sein, das GENOMatch-Konzept den Probanden/Patienten verständlich darzustellen und seine Ausgewogenheit zwischen der Garantie von Patientenrechten einerseits und der Ermöglichung wissenschaftlicher Forschung in der Pharmakogenetik andererseits zu vermitteln.

Zukünftige auf den Bereich des technischen Datenschutz bezogene Arbeiten werden sich – u. a. auf der Basis einschlägiger Betriebserfahrungen – mit der allgemeinen Modellierung und Bewertung von Pseudonymisierungsverfahren beschäftigen, um dadurch dem Datenschutz ein Instrument an die Hand zu geben, mit dem die Qualität solcher Konzepte wie des hier

vorgestellten möglichst exakt bestimmt werden kann.

## Literatur

- [1] 62. Konferenz der Datenschutzbeauftragten des Bundes und der Länder: *Entwurf eines „Gesetzes zur Sicherung der Selbstbestimmung bei genetischen Untersuchungen“*. 24.–26. Oktober 2001, <http://www.datenschutz-berlin.de/doc/de/konf/62/anlage.htm>
- [2] Council Of Europe: *Recommendation On The Protection Of Medical Data*. Recommendation No. R (97) 5 of the Committee of Ministers to Member States, adopted by the Committee of Ministers on 13 February 1997 at the 584th meeting of the Ministers' Deputies.
- [3] Metschke, R., Wellbrock, R.: *Datenschutz in Wissenschaft und Forschung*. Dez. 2002, <http://www.datenschutz-berlin.de>
- [4] Pfitzmann, A., Köhntopp, M.: *Anonymity, Unobservability, and Pseudonymity – A Proposal for Terminology*. In: H. Federath (Ed.): *Anonymity 2000*, Berlin u. a. (Springer, LNCS 2009) 2001, 1–9, and as Open Paper for Discussion on the Information Hiding Workshop 2001, Pittsburgh, PA, April 25-27, 2001.
- [5] Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein: *Kurzgutachten zum Datenschutzaudit „Konzept einer Datenverarbeitungsinfrastruktur der Fa. Schering AG für die sichere pseudonyme Einlagerung und Verwahrung von für genetische Analysen genutzten Blut- und Gewebeproben.“* <http://www.datenschutzzentrum.de/audit/k240603.htm>
- [6] UNESCO International Bioethics Committee: *Draft International Declaration On Human Genetic Data*. 28.8.2003. <http://unesdoc.unesco.org/images/0013/001312/131204e.pdf#page=27>