Blinde Anonymisierung: ein Verfahren für die Evaluation von Krebsfrüherkennungsprogrammen im Spannungsfeld zwischen Datenschutz und Datennutzung

Einleitung: Für die Evaluation der Prozess- und Ergebnisqualität von Krebsfrüherkennungsprogrammen (KFP) ist meist eine fallscharfe Verknüpfung medizinischer Daten aus verschiedenen Quellen notwendig. Dabei ist die Verarbeitung personenbeziehbarer Daten grundsätzlich nur erlaubt, wenn die betroffene Person eingewilligt hat oder eine Rechtsvorschrift dieses auf Grund eines wichtigen öffentlichen Interesses erlaubt. Während die Einholung von Einverständniserklärungen in den meisten Evaluationsstudien schon aufgrund der Größe der Kohorten impraktikabel ist, ist die Schaffung spezialgesetzlicher Regelungen, die eine Zusammenführung explizit für das jeweilige Vorhaben erlauben, meist aufgrund zeitlicher Beschränkungen keine Option. Daher können für solche Vorhaben nur bestehende Gesetze als Rechtsgrundlage herangezogen werden. Häufig wird darin gefordert, dass personenbezogene Daten nach Möglichkeit anonymisiert oder zumindest pseudonymisiert verarbeitet werden sollen sobald das Vorhaben dies gestattet.

2010 initiierte das Bundesamt für Strahlenschutz (BfS) eine Machbarkeitsstudie zur Mortalitätsevaluation des deutschen Mammographie-Screening-Programms (MSP) mit mehr als 10 Millionen anspruchsberechtigten Frauen (Förderkennzeichen: 3610S40002). Im Rahmen dieses Projekts entwickelten wir ein Verfahren, das eine pseudonymisierte Zusammenführung sowie anonymisierte Auswertung der Daten unter den gegebenen datenschutzrechtlichen Regelungen ermöglicht.

Material und Methoden: Unser Modell ist eine Weiterentwicklung der erprobten und evaluierten Verfahren zur Verknüpfung medizinischer Datensätze auf Basis deterministisch verschlüsselter Identitätsmerkmale, die das Epidemiologische Krebsregister NRW (EKR-NRW) seit 2005 eingesetzt hat. Da auch die für eine Evaluation benötigten Merkmale sogenannte Quasi-Identifikatoren bilden können, wurde dieses Konzept für die Anwendung in KFP modifiziert und um ein Anonymisierungsverfahren auf Basis verschlüsselter Auswertungsmerkmale (Blinde Anonymisierung) erweitert.

Das Konzept definiert den Datenfluss von den verschiedenen Datenhaltern über einen Pseudonymisierungsdienst (PSD) und eine Datenzusammenführende Stelle (DZS) an eine Evaluierende Stelle (ES). Bevor die Daten von den verschiedenen Datenhaltern an die DZS übertragen werden, findet eine umfangreiche Vorverarbeitung mithilfe einer Meldesoftware noch auf dem Rechner des jeweiligen Datenhalters statt. Die direkt personenidentifizierenden Daten (IDAT) werden analog zu den Verfahren des EKR-NRW noch auf dem Rechner des Melders deterministisch und irreversibel zu sog. Personenkryptogrammen (PKG) verschlüsselt. Die PKG dienen nach Überverschlüsselung durch den PSD, der fallscharfen Verknüpfung von Meldungen verschiedener Datenhalter in der DZS.

Für die medizinisch-epidemiologischen Auswertungsmerkmale (EDAT) werden ebenfalls bereits auf Seiten der Datenhalter sinnvolle Aggregationsstufen (z.B. PLZ: 5,4,3-stellig, oder Datumsangaben: tages-, monats-, quartals- und jahres-scharf, Kategoriebildung

medizinischer Merkmale) erzeugt. Danach wird jede Stufe mit einem deterministischen Verfahren so verschlüsselt, dass erst die ES die Daten wieder entschlüsseln kann. Die DZS erhält also keinerlei Klartextdaten.

Die so aufbereiteten Daten werden über den PSD (der die PKG noch einmal überverschlüsselt) an die DZS übertragen. Danach erfolgt mithilfe des stochastischen Record-Linkage die Zusammenführung der Teilmeldungen auf Basis der überverschlüsselten PKG, analog zum Verfahren des EKR-NRW. In regelmäßigen Intervallen stellt die DZS der ES einen anonymen Auswertungsdatensatz zur Verfügung. Als Anonymitätskriterium wurde k-Anonymität ausgewählt, da hierfür lediglich die Information benötigt wird, ob zwei Merkmalausprägungen gleich oder ungleich sind. Dies kann auch auf den aggregierten und deterministisch verschlüsselten EDAT Merkmalen geprüft werden. Die DZS wählt also für die quasi-identifizierenden EDAT Merkmale diejenigen Aggregationsstufen aus, mit denen der Datensatz dem k-Anonymitätskriterium genügt. Für die Quasi-Identifikatoren werden nur die so ermittelten und noch verschlüsselten Aggregationsstufen an die ES verschickt. Erst die ES kann diese wieder entschlüsseln und verfügt dann über den geprüft anonymen Auswertungsdatensatz. Datenschutzrechtliche Restriktionen greifen nun nicht mehr.

Ergebnisse: Das Konzept wurde für die Anwendung in der Modellregion NRW durch die zuständigen Datenschutzbehörden (LDI NRW, BfDI, BSI) geprüft und genehmigt. Es wurde ein „Proof-Of-Concept“ Prototyp entwickelt und erfolgreich an einem Simulationsdatensatz mit mehr als 1,5 Mio. Meldungen getestet. Die Simulationsdaten bestanden dabei aus realistisch verteilten IDAT für Frauen der Alterskategorie 50-69 aus NRW, sowie aus einigen potentiell quasi-identifizierenden EDAT Merkmalen, wie Postleitzahl und Geburtsdatum. Nach Vorverarbeitung der IDAT und EDAT und anschließender Verknüpfung der Meldungen mithilfe des Record-Linkage wurden basierend auf den verschiedenen verschlüsselten Aggregationsstufen unterschiedliche Generalisierungen (Selektionen von Aggregationsstufen) gebildet und anhand des k-Anonymitätskriteriums bewertet. Es zeigte sich, dass z.B. bei k=5 für eine 4-stellige Postleitzahl und ein monatsscharfes Geburtsdatum bereits eine Unterdrückung von über 2,5% notwendig wäre. Wählt man stattdessen nur bei einem der beiden Merkmale eine höhere Aggregationsstufen (also: eine 3-stellige Postleitzahl oder ein nur quartalsscharfes Geburtsdatum), liegt die notwendige Unterdrückungsrate für k=5 bei unter 0,1%. Mit diesen Generalisierungen würde selbst bei k=10 nur eine Unterdrückung von unter 1% notwendig sein.

Die wesentlichen Einflussfaktoren sind vor allem die Anzahl der Quasi-Identifizierenden Merkmale, die gewählten Aggregationsstufen, sowie der Umfang des Datensatzes. In einer konkreten Anwendung, wie der aktuell geplanten Evaluation der Brustkrebsmortalität im deutschen MSP, können in den EDAT weitere Quasi-Identifikatoren enthalten sein, die zu deutlich schlechteren Ergebnissen führen können. Der Vorteil des hier vorgestellten Ansatzes liegt jedoch darin, dass alle EDAT Merkmale wie Quasi-Identifikatoren verarbeitet werden, und somit jede mögliche Generalisierung erprobt werden kann. So können den zuständigen Datenschützern mit belastbaren Zahlen diejenigen Generalisierungen vorgeschlagen werden, die einerseits die geforderten Anonymitätskriterien erfüllen (k-Werte) und andererseits die für die

Auswertung notwendige Spezifität in den Merkmalen aufweisen. Unser Ansatz erlaubt also eine flexible Balance zwischen Informationstiefe und Schutz der Daten.

Aktuell wird die eigentliche Software implementiert und in der Modellregion NRW erprobt. Das Meldetool (SecuNym-RT) wurde bereits bei einem Datenhalter installiert und es konnten über 80.000 aus Echtdaten abgeleitete Testdatensätze erfolgreich verarbeitet und übertragen werden.

Diskussion: Eine Voraussetzung für unseren Ansatz ist eine hohe Qualität der Primärdaten, da die Datenaufbereitung hauptsächlich bei den Datenhaltern erfolgt. Dies unterstützt das Meldetool durch automatisierte Plausibilitätsprüfungen und eine Maske für manuelle Nachbearbeitungen.

Die größte technische Limitierung des Ansatzes besteht derzeit noch in der Anzahl der Datenhalter. Da alle Datenhalter für die Vorverarbeitung der Daten über dieselben geheimen Schlüssel verfügen müssen, muss ihre Anzahl limitiert werden.

Obwohl k-Anonymität ein vergleichsweise schwaches Anonymitätsmaß ist, setzen wir es aufgrund seiner Eigenschaft ein, auch auf verschlüsselten Daten prüfen zu können. Da dies zum Teil auch auf das stärkere Anonymitätsmaß l-Diversität zutrifft, prüfen wir derzeit eine entsprechende Erweiterung der Anonymitätsprüfung.

Derzeit müssen noch ca. 5% aller eingehenden Meldungen nach dem Record-Linkage in eine manuelle Nachbearbeitung der Zuordnung. Aktuell arbeiten wir an Verfahren, um diesen Anteil zu reduzieren.

Bisher müssen Generalisierungen in der der DZS manuell konfiguriert und akzeptable Unterdrückungsraten explorativ gefunden werden. Derzeit prüfen wir verschiedene Algorithmen für eine automatisierte Suche nach optimalen Generalisierungen.