**Supplemental Digital Content 1**

Sharing ICU patient data responsibly under the SCCM/ESICM Joint Data Science Collaboration: the AmsterdamUMCdb example

Patrick J. Thoral, Jan M. Peppink, Ronald H. Driessen, Eric J.G. Sijbrands, Erwin J.O. Kompanje, Lewis Kaplan, Heatherlee Bailey, Jozef Kesecioglu, Maurizio Cecconi, Matthew Churpek, Gilles Clermont, Mihaela van der Schaar, Ari Ercole, Armand R.J. Girbes, Paul W.G. Elbers on behalf of the *AmsterdamUMCdb Collaborators* and the *SCCM/ESICM Joint Data Science Task Force*

Contents

[eMethods 2](#_Toc52833258)

[eTable 1 General de-identification principles 3](#_Toc52833259)

[eTable 2 De-identification methods used in database creation 4](#_Toc52833260)

[eTable 3 Assumed background knowledge and assessment of re-identification risk after de-identification according to the HIPAA Privacy Rule 6](#_Toc52833261)

[References 7](#_Toc52833262)

# eMethods

The AmsterdamUMCdb data lake was created from the departmental Patient Data Management System (MetaVision, iMDsoft, Tel Aviv, Israel), the hospital wide EHR (Epic, Epic Systems, Verona, USA), the laboratory information system (GLIMS, MIPS, Ghent, Belgium), two legacy Microsoft Access disease severity and logistics databases developed locally, and a commercial Patient Scoring Database (MediScore, Itémedical, Tiel, The Netherlands).

All patient entries containing direct identifiers, including but not limited to names, addresses, telephone numbers and social security numbers, were removed. To additionally protect the privacy of healthcare providers, all information related to departments and individual healthcare providers was also removed. All dates and times were de-identified by transforming them to milliseconds after the first admission. We generalized year of admission, age at admission, weight and height, into bins with cut-offs for outliers. Free text fields were removed except one which accommodates specific laboratory results. This implies that no clinical notes have currently been released. For de-identification of the lab free text field, we used iterations of automated and manual techniques (1). All free text fields were manually checked by at least three authors.

# eTable 1 General de-identification principles

|  |  |  |
| --- | --- | --- |
| Term | Definition | Example |
| Masking | Hiding part of field or a complete field by random or pre-defined data | * Names: ‘Results discussed with Dr. XXXX’ * Dates: ‘Sample drawn on <<DATE REMOVED>>’ |
| Suppression | Removing a complete field or health record | * Social security numbers * Phone numbers * Patients with unique time series (i.e. high number of ICU readmissions) |
| Generalization | Reducing the precision of the data | * Age range (’60-70’) instead of actual age (‘63’) * User group (‘ICU nurse’) instead of actual user |

# eTable 2 De-identification methods used in database creation

| **Safe Harbor method:**  **removal of specific elements** | **Additional risk-based anonymization techniques** | **Comments** |
| --- | --- | --- |
| Names | Names of health care providers and departments: removed |  |
| All geographic subdivisions smaller than a state | All geographic data: removed |  |
| All elements of dates (except year) | Generalization to ranges:  admission year: 2003-2009, 2010-2016  age at (first) admission:  18-39, 40-49, 50-59, 60-69, 70-79, 80+  Measurements:  time relative to first admission |  |
| Telephone numbers |  |  |
| Vehicle identifiers and serial numbers, including license plate numbers |  | Not in source data |
| Fax numbers |  | Not in source data |
| Device identifiers and serial numbers |  | Not in source data |
| Email addresses |  |  |
| Web Universal Resource Locators (URLs) |  | Not in source data |
| Social security numbers |  |  |
| Internet Protocol (IP) addresses |  | Not in source data |
| Medical record numbers |  |  |
| Biometric identifiers, including finger and voice prints |  | Not in source data |
| Health plan beneficiary numbers |  |  |
| Full-face photographs and any comparable images |  | Not in source data |
| Account numbers |  |  |
| Any other unique identifying number, characteristic, or code, except as permitted by paragraph (c) of the Privacy Rule | Removal of table mapping admission IDs with random IDs during database creation |  |
|  | Height, generalized to ranges: 59-, 60-69, 70-79, 80-89,90-99, 100-109, 110+ | Suppressed when combination of *quasi-identifiers* uniquely identifies patients (*k*=1):   * 404 admissions (1.75%) |
|  | Weight, generalized to ranges:  159-, 160-169, 170-179, 180-189, 190+ | Suppressed when combination of *quasi-identifiers* uniquely identifies patients (*k*=1):   * 404 admissions (1.75%) |
|  | Diagnostic codes/categories: suppressed | Suppressed when combination of *quasi-identifiers* leads to the same or very similar diagnoses:   * 12 admissions (0.05%) |
|  | Complete patient or admission record when combination of *quasi-identifiers including suppression of quasi-identifiers* uniquely identifies patients (*k*=1):   * 60 patients (0.30%) * 265 admissions (1.13%) |  |

# eTable 3 Assumed background knowledge and assessment of re-identification risk after de-identification according to the HIPAA Privacy Rule

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **Hypothetical Adversary** | | |
|  |  | Friendly Researcher | Rogue  Researcher | Rogue  Insurance Company |
| **Assumed Background Knowledge** | |  |  |  |
|  | Gender | X | X | X |
|  | Age | X | X | X |
|  | Weight | X | X |  |
|  | Height | X | X |  |
|  | Admission date | X | X | X |
|  | Survival at discharge | X | X | X |
|  | Number of ICU admissions |  |  | X |
| **Assessment of re-identification risk** | |  |  |  |
|  | P(access) | 1.00 | 1.00 | 0.27 |
|  | P(intention) | 0.20 | 0.10 | 0.10 |
| Average Risk | |  |  |  |
|  | P(re-identification) | 0.726 | 0.726 | 0.205 |
|  | *k*-anonymity | 2 | 2 | 14 |
|  | *l*-diversity | 2 | 2 | 9 |
| Maximum Risk | |  |  |  |
|  | P(re-identification) | 1 | 1 | 1 |
|  | *k*-anonymity | 1 | 1 | 1 |
|  | *l*-diversity | 1 | 1 | 1 |
| **P final risk** | | 0.14 | 0.10 | 0.0055 |
| **Patients identified (*k*=1)** | | 12153 | 12153 | 1258 |

Average risk is used for determining final risk for the *friendly researcher* and the *rogue insurance company*, whereas maximum risk was used for the *rogue researcher*. For the *friendly researcher* P(intention) is acquaintance risk, the risk of knowing somebody in the database. Please note that calculation of *strict* average risk, requiring *k* of at least 2 for every combination of background knowledge, used in the risk-based de-identification in Table 2 is not available for the Safe Harbor method, since *k* = 1for many combinations of background knowledge.

# References

1. Neamatullah I, Douglass MM, Lehman L-WH, et al.: Automated de-identification of free-text medical records. *BMC Med Inform Decis Mak* 2008; 8:32