Jurnal Internasional Informatika Medis 126 (2019) 72–81

Daftar isi tersedia di SainsLangsung

Jurnal Internasional Informatika Medis

beranda jurnal: www.elsevier.com/locate/ijmedinf

# Proses ETL yang meningkatkan privasi untuk data biomedis

Fabian Prasser[1,*], Helmut Spengler[1], Raffael Bild, Johanna Eicher, Klaus A. Kuhn

Institut Informatika Medis, Statistik dan Epidemiologi, Rumah Sakit Universitas rechts der Isar, Universitas Teknik Munich, Ismaninger Str. 22, 81675 München, Jerman

INFO ARTIKEL

Kata kunci:
Pergudangan data klinis
Ekstrak Transform Load
Pribadi
Anonimisasi

ABSTRAK

Latar Belakang: Pendekatan modern berbasis data untuk penelitian medis memerlukan informasi tingkat pasien secara mendalam dan luas. Untuk membuat kumpulan data besar yang diperlukan, informasi dari sumber yang berbeda dapat diintegrasikan ke dalam gudang klinis dan translasi. Ini biasanya diterapkan dengan proses Ekstrak, Transformasi, Muat (ETL), yang mengakses, menyelaraskan, dan mengunggah data ke dalam platform analitik.

Objektif: Perlindungan privasi memerlukan pertimbangan yang cermat saat data dikumpulkan atau digunakan kembali untuk tujuan sekunder, dan anonimisasi data merupakan mekanisme perlindungan yang penting. Namun, lingkungan ETL umum tidak mendukung anonimisasi, dan alat anonimisasi umum tidak dapat dengan mudah diintegrasikan ke dalam pekerjaan ETLfladuh. Tujuan dari pekerjaan yang dijelaskan dalam artikel ini adalah untuk menjembatani kesenjangan ini.

Metode: Tujuan desain utama kami adalah (1) untuk mendasarkan proses anonimisasi pada metodologi penilaian risiko tingkat ahli, (2) untuk menggunakan metode transformasi yang menjaga kebenaran data dan sifat skematisnya (misalnya tipe data), (3) untuk mengimplementasikan metode yang mudah dipahami dan intuitif untuk dipahamifigambar, dan (4) untuk memberikan skalabilitas yang tinggi.

Hasil: Kami merancang sebuah novel dan eFFIproses anonimisasi yang efisien dan menerapkan plugin untuk platform Pentaho Data Integration (PDI), yang memungkinkan pengintegrasian anonimisasi data dan identifikasi ulangfianalisis risiko kation langsung ke pekerjaan ETLfladuh. Dengan menggabungkan different instance ke dalam proses ETL tunggal, data dapat dilindungi dari beberapa ancaman. Plugin ini mendukung kumpulan data yang sangat besar dengan memanfaatkan model pemrosesan berbasis streaming dari platform yang mendasarinya. Kami menyajikan hasil evaluasi eksperimental yang ekstensif dan mendiskusikan aplikasi yang berhasil.

Kesimpulan: Pekerjaan kami menunjukkan bahwa metodologi anonimisasi tingkat ahli dapat diintegrasikan ke dalam pekerjaan ETLfladuh. Implementasi kami tersedia di bawah lisensi open source non-restriktif dan mengatasi beberapa keterbatasan alat anonimisasi data lainnya.

## 1. Perkenalan

Penelitian medis modern membutuhkan data yang mendalam dan luas untuk meningkatkan pemahaman kita tentang perkembangan dan perjalanan penyakit dan pada akhirnya mengembangkan metode untuk pencegahan, diagnosis dan terapi yang ditargetkan. Dalam sistem kesehatan belajar“setiap pertemuan klinis berkontribusi pada penelitian dan penelitian diterapkan secara real time untuk perawatan klinis” [1]. Untuk menerapkan ini dalam skala besar, data harus dapat diakses, diselaraskan, dan terintegrasi[2,3]. Ini juga memerlukan penggunaan data untuk aplikasi sekunder yang melampaui tujuan pengumpulan awal[4,5].

Integrasi data dan khususnya gudang data adalah pusat dari effort. Dalam konteks ini, sistem basis data diatur yang mengintegrasikan data yang berbeda ke dalam tata letak umum yang efektifFFImendukung dengan baik analisis yang kompleks. Platform i2b2[6] adalah contoh terkenal dari sistem yang berfokus pada data yang dihasilkan oleh layanan klinis dan kesehatan dan oleh studi epidemiologi [7]. Platform terkait adalah transSMART, yang telah dikembangkan untuk analisis klinis dan integrated 'omics' data untuk penelitian translasi [8]. Beberapa institusi, seperti Vanderbilt University Medical Center[5], juga telah mengembangkan solusi khusus.

Data biasanya direplikasi dari sistem rutin ke gudang menggunakan proses ETL [9,10]: (1) datanya adalah diekstraksi dari sistem sumber, (2) dibersihkan, diselaraskan dan berubah ke dalam bentuk yang cocok untuk analisis, dan (3) sarat ke dalam solusi analitik. Untuk mengelola kompleksitas proses tersebut, mereka sering diimplementasikan menggunakanfic lingkungan, yangffer perpustakaan konektor untuk diffberbagai jenis sumber, operator transformasi, dan meja kerja grafis untuk

* Penulis yang sesuai.
Alamat email: fabian.prasser@tum.de (F. Prasser).
[1] Para penulis ini berkontribusi sama untuk pekerjaan ini.

https://doi.org/10.1016/j.ijmedinf.2019.03.006
Diterima 22 Juni 2018; Diterima dalam bentuk revisi 6 November 2018; Diterima 6 Maret 2019

menggabungkannya menjadi pekerjaan yang kompleksfladuh. Solusi terkenal adalah Pentaho Data Integration (PDI, juga dikenal sebagai Kettle)[11], yang merupakan alat standar untuk memuat data ke transSMART, dan Talend Open Studio (TOS) [12], yang merupakan komponen utama dari Integrated Data Repository Toolkit (IDRT) [13] untuk membuat gudang berbasis i2b2.

Saat mengumpulkan data medis atau saat menggunakannya kembali untuk tujuan sekunder, masalah privasi dan persyaratan hukum perlu dipertimbangkan dengan cermat. Perlindungan privasi melibatkan masalah etika, hukum dan sosial (ELSI) dan beberapa lapisan tindakan teknis dan non-teknis biasanya diperlukan untuk menerapkannya[14]. Di sisi teknis, privasi pasien dan pasien sering kali dilindungi oleh anonimisasi data, yang berarti bahwa kumpulan data diubah sedemikian rupa sehingga mencegah identifikasi ulang yang berhasil.fikation. Peraturan privasi nasional dan internasional menangani anonimisasi data. Di Amerika Serikat,Pelabuhan Aman
metode Aturan Privasi Undang-Undang Portabilitas dan Akuntabilitas Asuransi Kesehatan (HIPAA) menyediakan katalog atribut yang nilainya perlu dihapus atau dimodifikasified [15]. Selain itu,Penentuan Ahli memungkinkan penggunaan metode formal dan statistik untuk menilai dan mengelola identifikasi ulangfirisiko kation, yang mirip dengan cara penganoniman data perlu diterapkan di Uni Eropa [16].

Anonimisasi data adalah proses kompleks di mana pengurangan yang dihasilkan dari identifikasi ulangfirisiko kation perlu diseimbangkan dengan pengurangan utilitas data [14,17]. Berbagai macam diffmodel dan metode yang ada untuk transformasi data, penilaian risiko dan estimasi utilitas telah diusulkan untuk mengatasi pertukaran ini. Untuk mengelola proses yang kompleks ini, sejumlah alat telah dikembangkan, termasuk sdcMicro
[18], yang berfokus padaFFIstatistik sosial, dan ARX [19], yang memiliki spesifikasifi dirancang untuk aplikasi data biomedis dengan menerapkan metode yang telah direkomendasikan dalam fitua
[19-21]. Kedua alatffer tingkat kedewasaan yang tinggi dan mereka telah dimasukkan ke dalam oFFIpedoman cial, misalnya dari Badan Uni Eropa untuk Jaringan dan Keamanan Informasi (ENISA) [22] dan Badan Obat Eropa (EMA) [23].

### 1.1. Tujuan dan garis besar

Melakukan anonimisasi data dan identifikasi ulangfianalisis risiko kation sebagai bagian dari pekerjaan ETLflarus adalah persyaratan umum (lihat Bagian 4.1). Skenario aplikasi yang umum termasuk pemuatan data ke gudang klinis dan translasi, ekstraksi data dari pendaftar penelitian lintas institusi, dan berbagi data dengan kelompok penelitian eksternal. Namun, platform ETL seperti TOS atau PDI tidak menyediakan modul yang mendukung metode formal anonimisasi data dan identifikasi ulang.fianalisis risiko kation. Meskipun alat anonimisasi seperti sdcMicro atau ARX dapat digunakan untuk tujuan ini, mereka didasarkan pada lingkungan kerja mereka sendiri yang tidak dapat dengan mudah diintegrasikan ke dalam platform ETL (lihat Bagian 2).

Untuk menjembatani kesenjangan ini, kami telah mengembangkan plugin untuk platform ETL, yang mendukung anonimisasi data dan identifikasi ulang.fi penilaian risiko kation. Tujuan desain yang paling penting adalah (1) untuk memanfaatkan metodologi penilaian risiko tingkat ahli, (2) untuk menerapkan metode transformasi data yang menjaga kebenaran data input dan sifat skematisnya (misalnya tipe data), (3) untuk memanfaatkan proses anonimisasi yang mudah dipahami dan intuitif untuk dipahamifigambar, dan (4) untuk mencapai skalabilitas yang tinggi.

Untuk memenuhi tujuan desain ini, kami harus mengatasi berbagai tantangan. Pertama, kami perlu memutuskan desain dan lingkungan eksekusi yang sesuai untuk proses ETL. Kedua, kami perlu memilih dan mengintegrasikan metodologi untuk penilaian risiko dan anonimisasi yang terkenal,flfleksibel dan mudah dipahami. Ini melibatkan pengelolaan interaksi kompleks metode untuk mengukur dan mengurangi risiko privasi. Akhirnya, kami harus mengembangkan eFFIimplementasi yang efisien.

Sisa dari makalah ini disusun sebagai berikut: di Bagian 2 kami menjelaskan metode untuk penilaian risiko yang kami bangun, menyajikan metode anonimisasi baru, dan menjelaskan bagaimana kami telah menerapkannya. dan mengintegrasikannya ke dalam platform ETL yang ada. Di Bagian 3 kami menjelaskan bagaimana kami merancang eksperimen kami dan menyajikan hasilnya. Dalam Bagian 4 kita membahas hasil utama, aplikasi dalam praktek, dan melakukan perbandingan konseptual dengan pekerjaan sebelumnya. Di Bagian 5 kami menyimpulkan dan menunjukkan arah untuk pekerjaan di masa depan.

## 2. Bahan-bahan dan metode-metode

Metode kami untuk mengintegrasikan anonimisasi data dan penilaian risiko ke dalam proses ETL didasarkan pada metode yang telah ditetapkan untuk memperkirakan reidentifirisiko kation data medis, yang kami sajikan dalam fi bagian pertama dari bagian ini. Pada bagian kedua kami menyajikan algoritma anonimisasi baru yang telah kami kembangkan untuk memfasilitasi effintegrasi efektif dari metode ini ke dalam platform ETL. Bagian terakhir dari bagian ini berfokus pada bagaimana kami menerapkan metode ini dan bagaimana kami mengintegrasikannya ke dalam platform ETL yang konkret.

### 2.1. Model umum untuk penilaian risiko

Identifikasi ulangfikation adalah ancaman utama yang ditangani oleh undang-undang dan peraturan [15,16] dan model untuk mengukur risiko terkait sangat penting bagi anonimisasi data dan manajemen risiko privasi. Identifikasi ulangfi kation dapat dipahami sebagai proses keterkaitan [24]: keunikan (kombinasi) atribut dieksploitasi untuk menghubungkan catatan kumpulan data dengan data tambahan atau pengetahuan latar belakang musuh. Atribut yang dapat digunakan untuk membuat tautan disebutkuasi-identifiers [25]. Contoh umum termasuk data demografi dan informasi lain yang mungkin diketahui musuh, seperti status pendidikan atau pekerjaan employment[21]. Menerapkan perlindungan perlu mempertimbangkan berbagai faktor, misalnya tujuan kemungkinan penyerang, kemampuan meniru dan membedakan data yang akan dilindungi, dan ketersediaan latar belakang pengetahuan[26,27]. Tiga diff Skenario ancaman yang ada dapat dibedakan [28]. Dibawah

itu jaksa model, musuh diasumsikan menargetkan target tertentufic individu dan untuk mengetahui bahwa data tentang individu ini terkandung dalam dataset. Risiko serangan yang berhasil dapat dihitung, berdasarkan pada pembedaan catatan dalam dataset mengenai quasi-identi-
fiers [26]. Namun, telah ditunjukkan bahwa metode ini signifikanfiterlalu melebih-lebihkan risiko dalam banyak kasus [29]. Di bawahwartawan model, musuh diasumsikan menargetkan individu yang sewenang-wenang tanpa pengetahuan sebelumnya tentang keanggotaan. Seringkali, pengetahuan latar belakang ini jauh lebih realistis daripada dalam model penuntut, karena kumpulan individu yang diwakili dalam kumpulan data hanyalah sampel dari populasi yang lebih besar. Namun, fakta bahwa pengetahuan tentang populasi biasanya tidak tersedia membuatnya juga berbedaFFIkultus untuk secara andal menentukan dan mengelola risiko serangan jurnalis yang berhasil. Akhirnya, di bawahpemasar model, musuh diasumsikan bertujuan untuk mengidentifikasi kembali sebanyak mungkin individu. Dengan demikian, risiko serangan yang berhasil dapat dinyatakan sebagai jumlah rata-rata yang diharapkan dari identifikasi ulangfied individu.

El Emam telah mengusulkan metodologi yang menggabungkan perkiraan risiko di bawah model yang sudah ada ini [28]. Karena risiko jurnalis tidak dapat diukur secara kuantitatiffied dalam banyak kasus, metodologi memanfaatkan fakta bahwa risiko jaksa selalu merupakan batas atas risiko jurnalis dan pemasar. Risiko jaksa adalah kuantitatiffied untuk semua catatan dan digabungkan menjadi tiga langkah global. Itufiukuran pertama adalah Risiko Tertinggi ($R_{h)}$. Ini kuantitatif fiAda risiko dalam skenario kasus terburuk, yaitu serangan jaksa
terhadap rekor dengan re-identi tertinggifirisiko kation di seluruh dataset. Untuk setiap catatanr, identifikasi ulangfirisiko kation dihitung sebagai $\frac{1}{f_r}$, *dari*
dimana $f_r$ adalah jumlah record dalam dataset yang tidak dapat dibedakan dari r tentang quasi-identifier (termasuk r diri).
Seperti disebutkan sebelumnya, ini juga merupakan batas atas risiko dalam skenario lain, yaitu untuk serangan jurnalis atau jaksa. Bahkan ketika risiko ini dibatasi oleh ambang batas, penyerang dapat berharap untuk mengidentifikasi kembali sejumlah individu tertentu melalui hubungan acak dengan catatan yang cocok. Ini adalah ditangkap oleh ukuran kedua, Risiko Rata-rata ($R_{Sebuah)}$, yang menyediakan

terikat lebih ketat untuk risiko pemasar. Untuk menjelaskan fakta bahwa model jaksa didasarkan pada asumsi kasus terburuk, langkah ketiga, dipanggil Catatan berisiko ($R_{f)}$ dapat digunakan untuk sedikit melonggarkan persyaratan perlindungan. Ini mengungkapkan frekuensi catatan yang terkait dengan identifikasi ulangfirisiko kation lebih tinggi dari ambang batas yang diberikan . De formalfipemahaman dari ketiga tindakan risiko ini disediakan di Bagian A dari panduan tambahan file.

Dengan metodologi ini dan hanya satu spesifikasi penggunafiparameter ed (), tiga ukuran risiko intuitif dapat diturunkan yang mengukur kerentanan data terhadap semua jenis serangan yang dipertimbangkan. Pada saat yang sama, model memfasilitasi keseimbangan perlindungan privasi dan kegunaan data, karena memungkinkan pengguna untuk mengizinkan bahwa sebagian kecil dari catatan memiliki risiko yang lebih tinggi dari ambang batas.. Diberikan suFFIkecil sekali θ dan suFFIsedikit arsip yang berisiko, tingkat perlindungan yang tinggi dapat diasumsikan, karena sangat kecil kemungkinan bahwa arsip yang ditargetkan dalam serangan (jaksa atau jurnalis) adalah salah satu arsip yang melebihi ambang batas [28]. ambang batas$\tau_{Sebuah}$ untuk risiko rata-rata $R_{Sebuah}$ dan $\tau_h$ Untuk risiko tertinggi $R_h$ dapat diperkenalkan selain θ untuk menentukan tingkat perlindungan yang harus memuaskanfied oleh prosedur anonimisasi data.

## 2.2. Metode anonimisasi baru

Secara otomatis mengubah data sedemikian rupa sehingga memenuhi spesifikasi penggunafiAmbang batas risiko ini rumit dan memerlukan integrasi model risiko dengan teknik dan metode transformasi data untuk mengukur utilitas data. Menghasilkan data keluaran yang benar menyiratkan bahwa data masukan tidak terganggu dan tidak ada data sintetis yang dihasilkan, yang sangat penting dalam penelitian medis di mana masuk akal dan kebenaran adalah pusat[30]. Oleh karena itu, kami memutuskan untuk tidak menggunakan skema transformasi yang menggunakan penambahan noise[31] atau kumpulan data [32]. Selain itu, kami ingin memastikan bahwa metode kami dapat diintegrasikan ke dalam pekerjaan ETL yang ada. flmengalir tanpa perlu memodifikasi representasi data perantara atau target. Ini menyiratkan bahwa sifat skematik dari data input harus dipertahankan, yang berarti bahwa tipe data tidak boleh diubah dan tidak ada atribut tambahan yang harus dimasukkan ke dalam tabel dan baris yang diproses. Jadi kami tidak dapat menggunakan generalisasi data[25] atau bucketisasi [33].

Berdasarkan pertimbangan ini, kami memutuskan untuk menerapkan algoritma penekanan sel. Dengan model ini, ambang batas risiko ditegakkan dengan menghapus nilai atribut individual dari catatan individual. Metode ini membutuhkan nol configurasi (selain menentukan ambang risiko), data keluaran adalah benar dan properti skematik dipertahankan. Selain itu, hasilnya sangat cocok untuk melakukan analisis statistik umum, asalkan effdll dari penekanan sel dipertimbangkan (misalnya dengan imputasi) [28,30,34].

Gambar 1 menunjukkan bagaimana penekanan sel dapat digunakan untuk melindungi dataset dari duaffskenario ancaman yang ada. Dalam sederhana inified contoh, kumpulan data klinis dilindungi dari serangan pemasar oleh penyerang eksternal menggunakan atribut demografis {Usia, Jenis Kelamin, Wilayah} dan dari serangan jaksa oleh penyerang internal menggunakan atribut klinis {Berat, ICD-10}. Nilai yang ditekan (yang dilambangkan dengan *) diperlakukan sebagai kategori sendiri, yang berarti nilai yang ditekan hanya dianggap sama dengan nilai yang ditekan lainnya. Di bawah asumsi ini semua set baris yang berisi nilai atribut pengidentifikasi kuasi yang sama adalah terputus-putus berpasangan dan membentuk apa yang disebutkelas kesetaraan.

Setiap kelas ekuivalensi menggambarkan satu set catatan yang tidak dapat dibedakan dengan penyerang dan karenanya ukurannya menentukan risiko keberhasilan identifikasi ulang.fikation. Dalam contoh, kelas kesetaraan diilustrasikan dengan garis putus-putus. Dengan menekan 20 dari 50 nilai atribut dalam kumpulan data (40%), risiko serangan eksternal yang berhasil turun dari 60% ($R_{a=10}$ $\frac{6}{}$) hingga 30% ($R_{a=10}$ $\frac{3}{}$) dan risiko internal yang sukses serangan turun dari 100% ($R_{h=1}$ $\frac{1}{}$) hingga 33% ($R_{h=3}$ $\frac{1}{}$). Contohnya juga menunjukkan bahwa penekanan sel menantang untuk menerapkan eFFIsecara efisien, karena ruang solusi potensial untuk kumpulan data yang diberikan terdiri dari $2^{(tidak\cdot m)}$ transformasi dimana tidak adalah jumlah record dan saya adalah jumlah atribut yang dapat digunakan untuk linkage. Ini sama dengan $2^{50}$ solusi potensial sudah dalam contoh sederhana kami. Jadi penekanan sel biasanya dilakukan dengan menggunakan algoritma heuristik.

Implementasi kami mengikuti pendekatan ini dengan menegakkan secara rekursif pengguna-defiambang batas $\tau_{Sebuah}$ dan $\tau_h$ untuk subset dari input dataset. Ini diimplementasikan dengan ARX, yang mampu menghitung optimal solusi untuk masalah anonimisasi data yang spesifikfied sebagai berikut [35]: (1) semua ambang batas risiko harus dipenuhi, (2) setiap kolom yang berisi nilai pengenal kuasi dapat disimpan apa adanya atau dihilangkan seluruhnya (penekanan atribut), (3) spesifikasi tertentufied jumlah catatan dapat sepenuhnya ditekan (disebut batas penekanan), (4) jumlah keseluruhan sel yang ditekan harus minimal. Metode kami menjalankan proses ini secara rekursif untuk record yang telah diredam, seperti yang diilustrasikan pada Gambar 2.. Dalam setiap iterasi,$\tau_h$ dan $\tau_{Sebuah}$ diberlakukan pada satu set catatan; yang lain ditekan. Kami menggunakank-model privasi anonimitas untuk melaksanakan $\tau_h$ [25] dan menegakkan $\tau_{Sebuah}$ dengan menentukan batas atas pada rata-rata aritmatika dari catatan' identifikasi ulangfirisiko kation. Tambahan parameter $aku_s$ spesifikasifies jumlah maksimum panggilan rekursif dengan de-fining batas penekanan untuk setiap iterasi. Pseudocode yang mengilustrasikan metode anonimisasi secara lebih rinci dan diskusi tentang implikasi untuk kualitas data disediakan di Bagian B dari pelengkap file. Sementara proses ini sangat eFFIilmiah dan effektif, seperti yang akan kami tunjukkan di bagian berikutnya, tetap perlu untuk menunjukkan bahwa itu benar-benar benar. Sangat mudah untuk melihat bahwa menegakkan ambang keseluruhan pada re-identifikasi tertinggi tifirisiko kation $R_h$ dapat dilakukan dengan menerapkan ambang yang sama pada subset record yang terpisah. Namun, tidak sepele untuk melihat bahwa ini that proses dapat digunakan untuk menerapkan ambang batas global pada rata-rata re-identitasfirisiko kation $R_{Sebuah.}$ Sebuah bukti disediakan di Bagian C dari pelengkap file.

## 2.3. Implementasi dan integrasi

Untuk membuat solusi kami dapat diakses oleh spektrum pengguna yang luas, kami

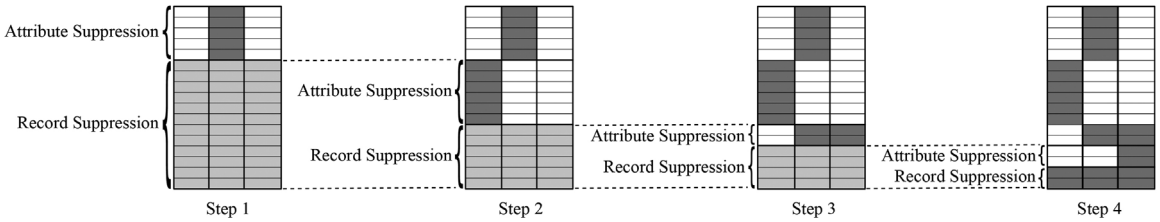
| Age | Sex | Region | Weight | ICD-10 |
|---|---|---|---|---|
| 53 | F | North | 73 | C18.7 |
| 68 | F | North | 73 | C18.7 |
| 68 | M | North | 82 | C18.7 |
| 68 | M | North | 77 | C18.7 |
| 71 | M | North | 73 | C18.2 |
| 71 | M | North | 67 | C18.2 |
| 68 | M | South | 67 | C18.2 |
| 68 | F | South | 67 | C18.7 |
| 68 | F | South | 67 | C18.7 |
| 68 | F | South | 67 | C18.7 |

(a) Input dataset

| Age | Sex | Region | Weight | ICD-10 |
|---|---|---|---|---|
| * | * | North | * | C18.7 |
| * | * | North | * | C18.7 |
| * | * | North | * | C18.7 |
| * | M | North | * | C18.7 |
| * | M | North | * | C18.2 |
| * | M | North | * | C18.2 |
| 68 | * | South | * | C18.2 |
| 68 | * | South | 67 | C18.7 |
| 68 | * | South | 67 | C18.7 |
| 68 | * | South | 67 | C18.7 |

(b) Output dataset

Gambar 1. Contoh dataset sebelum (a) dan setelah (b) telah ditransformasi menggunakan penekanan sel. Garis putus-putus menggambarkan kelas kesetaraan sehubungan dengan dua diffhimpunan kuasi-identifier: {Usia, Jenis Kelamin, Wilayah} dan {Berat, ICD-10}.




Gambar 2. Ilustrasi algoritma penekanan sel rekursif. Dalam setiap langkah rekursi, algoritma menentukan keseimbangan optimal antara atribut dan penindasan rekaman.

memutuskan strategi implementasi dua langkah. DalamfiLangkah pertama, anonimisasi yang dijelaskan dan metodologi penilaian risiko diimplementasikan ke dalam ARX. Ini memungkinkan kami untuk memanfaatkan kerangka kerja anonimisasi yang sangat skalabel[19] untuk membuat penilaian risiko dan operator anonimisasi yang kemudian dapat diintegrasikan ke dalam lingkungan ETL pada langkah kedua. Dalam konteks ini, kami memutuskan untuk mengembangkan plugin untuk platform PDI karena beberapa alasan. Pertama, kami sering menggunakan PDI untuk memuat data ke transSMART. Kedua, antarmuka yang disediakan oleh PDI cukup intuitif sementara kurva pembelajaran untuk TOS dapat dianggap agak lebih curam. Ketiga, PDIffmenyediakan serangkaian fitur yang luas dalam edisi komunitasnya (mis. penyebaran ke cluster) sementara fitur TOS yang paling canggih hanya tersedia melalui lisensi komersial. Selain itu, dengan rilis baru-baru ini (versi 8.0), antarmuka pemrograman platform PDI telah diterima secara signifikanfitidak bisa modernisasi.

Di meja kerja PDI, proses ETL dapat dimodelkan sebagai grafik terarah, di mana sumber data, transformasi, dan sink data direpresentasikan sebagai node yang disebut “Langkah”. Data flow antara node diwakili oleh tepi. Data yang tidak dapat diproses dapat dianotasi dengan informasi tambahan dan diarahkan ke keluaran kesalahan khusus. Dengan menggabungkan beberapa langkah, proses ETL kompleks yang mengintegrasikan sumber heterogen dapat dirancang, dijalankan, dan dipantau.Gambar 3 menunjukkan tangkapan layar dari proses ETL di mana data dari tigaffsumber data erent (sebuah CSV file, database relasional, dan aliran pesan HL7) digabungkan, divalidasi, diubah, dan fiakhirnya dimuat ke dalam database target.

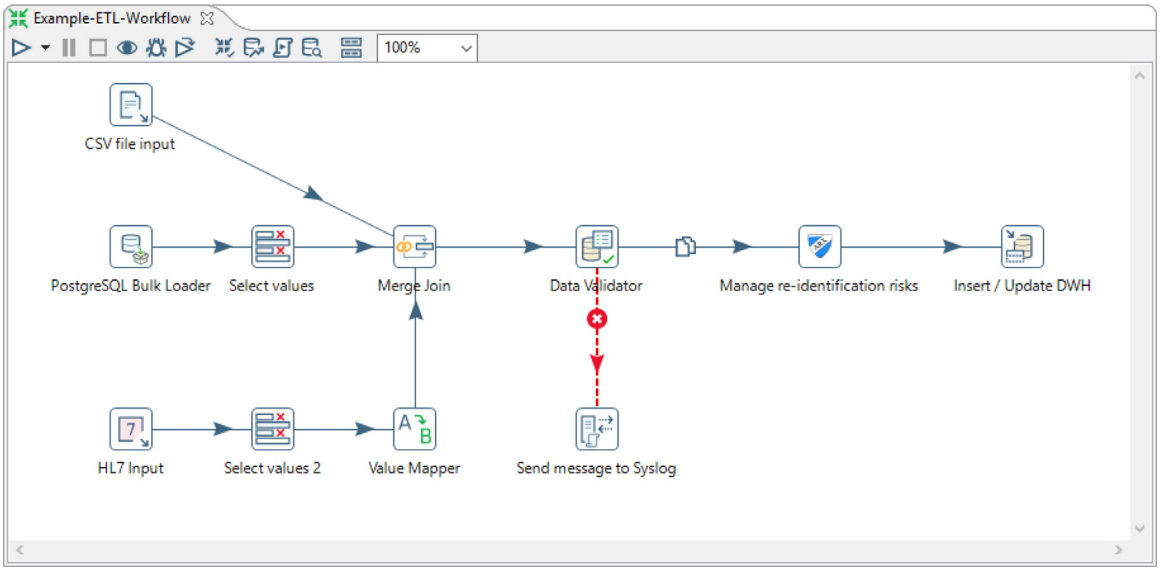
Pemrosesan data di PDI berorientasi pada aliran dengan satu baris data yang merupakan unit atomik dan terisolasi dari aliran data. Ini berarti bahwa data dilewatkan melalui jalur pipa ETL baris demi baris. Ini memungkinkanparalelisme pipa melintasi rantai langkah. Namun, ini juga menyiratkan bahwa plugin yang memerlukan tampilan holistik pada kumpulan data keseluruhan, seperti plugin kami untuk menilai risiko atau menganonimkan data, perlu diubah.ffeh baris yang masuk. Ada trade-offs terlibat dalam mengimplementasikan ini, karena yang terakhir mematahkan paralelisme pipa dan set data volume tinggi bisa terlalu besar untuk sepenuhnya mewujudkannya di memori utama.

Untuk mengatasi masalah ini, kami menerapkan teknik yang disebut pemblokiran baris. Ini berarti bahwa plugin kami mewujudkan kumpulan catatan (yaitu blok) dari de penggunafiukuran tertentu, yang kemudian dianalisis atau dianonimkan. Segera setelah setiap blok diproses, baris yang ada diteruskan ke plugin berikutnya dalam pekerjaanfladuh. Akibatnya, paralelisme dapat dipertahankan dan kumpulan data yang sangat besar dapat diproses. Dalam hal perlindungan privasi, pendekatan ini dijamin benar (lihat Bagian C dari suplemenfisaya).

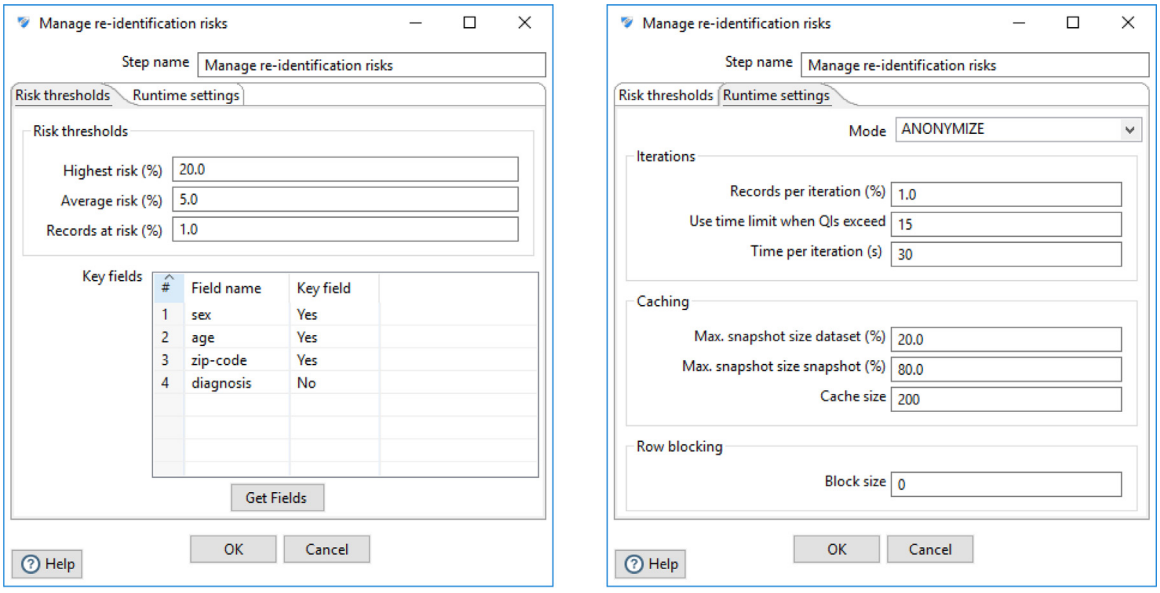
Kami menerapkan semua metode ke dalam plugin untuk platform PDI. Implementasi kami tersedia sebagai perangkat lunak sumber terbuka[36,37] yang kompatibel dengan versi terbaru 8.0 dari PDI. Plugin menyediakan metode untuk mengidentifikasi ulangfianalisis risiko kation dan anonimisasi data. Ini kompatibel dengan semua fungsi dan plugin PDI lainnya.

tab Ambang batas risiko, yang ditampilkan dalam Gambar 4(a), memungkinkan pengguna untuk tentukan kuasi-identifiers dan ambang batas yang dijelaskan sebelumnya. Kompatibel dengan model relasional yang mendasari lingkungan ETL, nilai-nilai yang ditekan diganti denganBATAL. Dengan demikian skema dan tipe data dari data input dipertahankan. Ketika risiko dinilai dan salah satu dari mereka melebihi pengguna-defiambang batas tertentu, data yang masuk tidak akan ditransfer ke langkah berikutnya dan, jika diinginkan, dapat dialihkan ke pintu keluar kesalahan. Langkah-langkah risiko dicetak ke konsol untuk tujuan logging. tabPengaturan waktu proses, yang ditampilkan dalam Gambar 4(b), dapat digunakan untuk menentukan parameter affmempengaruhi perilaku runtime dari algoritma anonimisasi.

Untuk mengatasi beberapa skenario ancaman, data dapat dilewatkan melalui berbagai caraffcontoh-contoh dari plugin confidiharapkan untuk mengatasi diffskenario ancaman yang ada (lih. contoh dalam Gambar 1). Ini dimungkinkan karena plugin mempertahankan properti skematik dari data input dan karena itu membuat




Gambar 3. Proses ETL khas di lingkungan desain PDI Spoon.




(a) Privacy related parameters. (b) Parameters determining runtime behavior.

Gambar 4. Tangkapan layar dari konfigurasi pluginfidialog gurasi.

penggunaan diffcara-cara baru untuk menafsirkan nilai-nilai yang ditekan. Selama anonimisasi, nilai-nilai yang ditekan diperlakukan sebagai kategori sendiri, yang berarti bahwaBATAL hanya cocok BATAL ketika menghitung pembedaan catatan. Namun, dalam rantai langkah anonimisasi dengantumpang tindih kuasi-identifier, ini dapat menyebabkan situasi, di mana satu operasi anonimisasi membatalkan jaminan privasi yang telah diberlakukan pada langkah sebelumnya karena kategori baru dimasukkan ke dalam variabel pengidentifikasi kuasi yang dibahas sebelumnya (contoh dapat ditemukan di Bagian D dari fisaya). Untuk alasan ini, ketika
menilai risiko, plugin kami menafsirkan nilai yang ditekan sebagai kartu liar. Ini berarti bahwa mereka dapat menandingi yang lain (ditekanatau tidak dimodifikasi fied) nilai, yang menghindari masalah ini. Meskipun telah ditunjukkan bahwa interpretasi ini dapat memberikan vektor serangan kepada musuh dalam keadaan yang jarang terjadi[38], kami menunjukkan bahwa ini adalah interpretasi standar dalam fibidang kontrol pengungkapan statistik dan juga default di sdcMicro.

## 3. Hasil

### 3.1. Pengaturan eksperimen

Di bagian ini, kami menyajikan hasil evaluasi skalabilitas solusi kami serta kualitas data keluaran, termasuk perbandingan dengan pekerjaan sebelumnya. Kami menunjukkan bahwa batasan teoretis pada kualitas data yang disediakan oleh pendekatan kami tidak dapat dengan mudah diperoleh (untuk diskusi tentang aspek optimalitas, kami merujuk ke Bagian B dari panduan tambahanfi saya). Oleh karena itu, kami fokus pada evaluasi eksperimental dengan kumpulan data dunia nyata untuk menganalisis bagaimana kinerja metode dalam praktik. Kami melakukan empat diffset eksperimen:

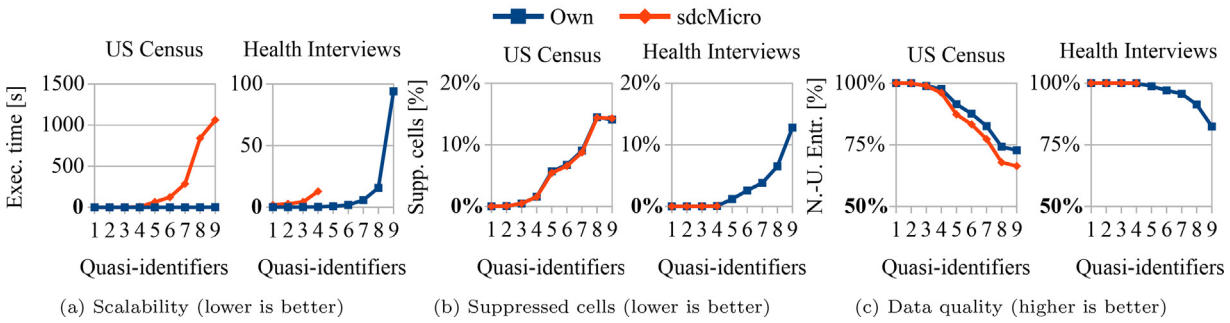
- Perbandingan dengan pekerjaan sebelumnya: Kita fipertama-tama bandingkan kinerja plugin kami dengan sdcMicro (versi 5.0.3) [18], yang memiliki sel algoritma supresi yang telah diimplementasikan dalam C++ dan ditautkan ke dalam perangkat lunak. Selanjutnya, kami mempelajari kegunaan data keluaran yang dihasilkan oleh metode penekanan sel kami dibandingkan dengan metode transformasi data lainnya menggunakan konsep perlindungan privasi. ving kubus data yang diusulkan oleh Kim et al. [39].
- Perbandingan menggunakan diffskenario ancaman saat ini: sdcMicro dan karya Kim et al. fokus pada skenario ancaman sederhana, sementara aplikasi kami proach mendukung kombinasi dari beberapa diffambang risiko yang ada. Kami melakukan eksperimen tambahan menggunakan berbagai parameterisasi dan kualitas data keluaran terukur untuk mempelajari effdll.
- Analisis trade-o risiko-utilitasffs: Di set ketiga percobaan kami membangun batas utilitas risiko, yang merupakan plot yang memvisualisasikan tukar tambahffs bahwa metode anonimisasi menyediakan antara privasi perlindungan dan kualitas data [40].
- Analisis effdll dari pemblokiran baris: Parameter yang di- cifies ukuran blok memiliki berbagai dalamflmempengaruhi kualitas output data dan waktu eksekusi proses anonimisasi. Di sebuahfirangkaian percobaan terakhir kami mempelajari e . iniffuntuk menentukan apakah pemblokiran baris adalah efekffmekanisme efektif untuk memproses kumpulan data besar dengan plugin kami.

Kami menggunakan dua dataset, yang berbedaffer dalam ruang lingkup dan ukuran dan yang memiliki sudah digunakan untuk mengevaluasi pekerjaan sebelumnya tentang anonimisasi data: (1) Sensus AS, kutipan dari 30.162 catatan dari database sensus 1994, yang berfungsi sebagai standar de-facto untuk evaluasi algoritma anonimisasi, dan (2) Wawancara Kesehatan, satu set 1.193.504 tanggapan terhadap survei kesehatan besar. Untuk penjelasan rinci kita lihat[41]. Untuk setiap dataset kami memilih hingga sembilan quasi-identi- fiers, yang terdiri dari data demografis dan atribut lebih lanjut, yang sering dianggap terkait dengan risiko tinggi identifikasi ulangfikation [21]. Semua percobaan dilakukan pada mesin desktop yang dilengkapi dengan CPU Intel Core i5 quad-core 3,2 GHz yang menjalankan sistem operasi Windows 7 64-bit. Platform PDI (versi 8.0) dijalankan menggunakan Oracle JVM 64-bit (1.8). Jumlah iterasi yang dilakukan oleh
algoritma (parameter $aku_{s)}$ ditetapkan ke 100 di semua percobaan.

### 3.2. Perbandingan eksperimental dengan pekerjaan sebelumnya

Kita fipertama bandingkan plugin kami dengan sdcMicro [18]. Penekanan sel algoritma sdcMicro telah diimplementasikan dalam C++ dan ditautkan ke dalam paket untuk meningkatkan skalabilitas. Perangkat lunak ini hanya mendukung penekanan sel untuk menegakkan ambang batas pada risiko tertinggi. Oleh karena itu kami menetapkan $\tau_{r}$ (catatan berisiko) menjadi nol dan menggunakan ambang batas pada penuntutan ulang identitasfirisiko kation ($\tau_{h)}$ 20%, yang merupakan parameterisasi umum common [21].

Gambar 5(a) menunjukkan waktu eksekusi yang diukur sambil meningkatkan jumlah atribut quasi-identifying. Dapat diamati bahwa implementasi kami signifikanfijauh lebih terukur daripada sdcMicro. Meskipun metode kami dapat dengan mudah menangani kumpulan data Sensus AS terlepas dari




Gambar 5. Perbandingan hasil yang diperoleh dengan plugin kami dan hasil yang diperoleh menggunakan sdcMicro. Kami melaporkan waktu eksekusi rata-rata, jumlah yang ditekan sel dan kuantitas kualitas datafied dengan model Entropi Non-Seragam.

jumlah quasi-identifidipilih (≤2 s di semua konfigurations), sdcMicro sudah membutuhkan lebih dari 1000 s untuk memproses dataset dengan sembilan quasi-identifier confidiperkirakan. Selanjutnya, sdcMicro tidak dapat menangani dataset Wawancara Kesehatan dalam 1800 detik ketika lebih dari empat identifikasi kuasifiers itu spesifikfied. Untuk alasan praktis, kami membatalkan semua eksperimen menggunakan sdcMicro yang tidak selesai dalam jangka waktu ini. Plugin kami umumnya memproses dataset ini dalam waktu tidak lebih dari 94 detik. Dapat dilihat bahwa kedua implementasi adalahffdipengaruhi oleh peningkatan eksponensial dalam ukuran ruang solusi dengan peningkatan jumlah kuasi-identifiers [35]. Namun, plugin kami dapat dikonfidiperkirakan menggunakan effalgoritma heuristik efektif ketika ruang solusi menjadi terlalu besar [42].

Mengenai kualitas data, kami mengukur jumlah sel yang sebanding yang ditekan oleh metode kami dan oleh sdcMicro (Gambar 5(b)). Akhirnya,Gambar 5(c) menunjukkan bagaimana anonimisasi berdampak pada distribusi nilai atribut. Untuk mengukur ini, kami menggunakan model Entropi Non-Seragam[43] yang sering digunakan untuk menilai kualitas de-identified data dan didasarkan pada konsep saling informasi [28]. Kami menormalkan hasil yang diperoleh oleh model ini sedemikian rupa sehingga 100% mewakili dataset input asli sementara 0% mewakili dataset dari mana semua nilai telah dihapus. Terlihat bahwa kualitas data menurun ketika jumlah quasi-identifiers meningkat, terutama untuk dataset yang lebih kecil Sensus AS. Dapat juga diamati bahwa metode kami memiliki dampak yang lebih kecil pada distribusi nilai atribut, yang menyiratkan penerapan penekanan nilai yang lebih seimbang.

Baru-baru ini, Kim et al. melakukan evaluasi eksperimental effdll dari different metode anonimisasi data saat menerapkan gudang perlindungan privasi untuk data medis [39]. Dalam studi mereka, data dianonimkan dan kemudian dikumpulkan ke dalam kubus data, yang merupakan model yang digunakan dalam aplikasi pergudangan. Penulis kemudian mengukur hilangnya informasi yang disebabkan oleh metode anonimisasi dan ketepatan hasil dari dua jenis kueri yang dikeluarkan terhadap kubus data:pertanyaan titik, yang menghitung jumlah record yang cocok dengan spesifikasi tertentufic kombinasi nilai atribut dan kueri rentang, yang menghitung jumlah record yang cocok dengan kombinasi rentang di atas domain nilai atribut. Mereka mempelajari dua pendekatan berbasis generalisasi dan satu algoritma bucketization.

Kami dengan tepat mereproduksi pengaturan eksperimental mereka, yang juga menggunakan kumpulan data Sensus AS, dan membandingkan hasil yang diperoleh menggunakan metode kami dengan hasil yang disajikan dalam [39]. Untuk spesifikasi yang tepat fikation algoritme dan diskusi mendalam tentang hasil yang kami rujuk ke Bagian E pelengkap file. Seperti yang bisa dilihat diTabel 1, metode kami mengungguli kedua pendekatan berbasis generalisasi dalam hal kehilangan informasi, berkinerja sangat baik pada kueri titik dan memberikan kinerja yang wajar pada kueri jangkauan. Pada saat yang sama, metode kami adalah satu-satunya pendekatan yang dipertimbangkan dalam eksperimen yang mempertahankan sifat skematik dari data input, dan jauh lebih mudah untukfidaripada algoritma berbasis generalisasi.

### 3.3. Analisis eksperimental menggunakan di usingffskenario ancaman saat ini

Plugin kami mendukung ambang batas pada identifikasi ulang jaksafirisiko kation ($\tau_{Th}$) dan mengidentifikasi ulang pemasarfirisiko kation ($\tau_{Sebuah}$). Risiko rata-rata ketat [21] adalah model privasi umum yang menggabungkan kedua ambang risiko. Untuk menganalisis perbaikan utilitas data yang dapat diperoleh dengan menggunakan model ini, kami telah melakukan perbandingan kedua pendekatan. Sebagai ambang batas risiko, kami juga menggunakan 20%. Kami menggunakan ambang yang sama sekali untuk mengendalikan risiko jaksa dan sekali untuk mengendalikan identifikasi ulang pemasarfi risiko kation tetapi menggabungkan yang terakhir dengan ambang 50% pada risiko jaksa, yang memastikan bahwa tidak ada catatan yang diidentifikasi secara unikfi sanggup. Kami mencatat bahwa perbandingan ini hanya berfokus pada plugin kami, karena risiko rata-rata yang ketat sepengetahuan kami tidak didukung oleh alat lain.

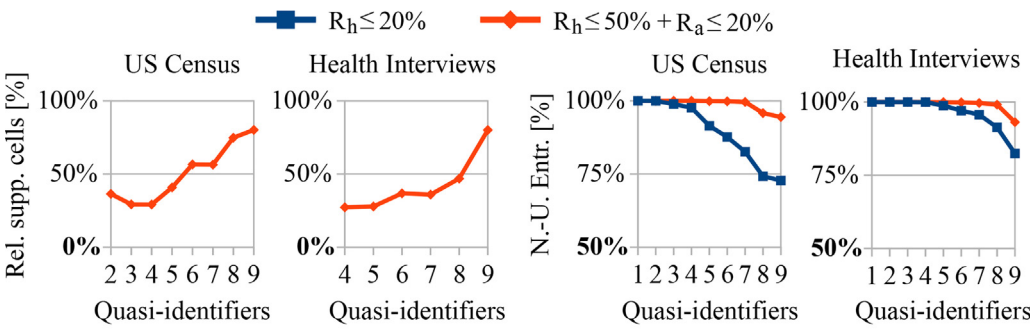
Kami mengukur tidak ada signifikansifitidak bisafferences dalam waktu eksekusi saat menggunakan dua model. Namun, kami mengamati peningkatan penting dalam kualitas data saat menggunakan risiko rata-rata yang ketat. Gambar 6(a) menunjukkan jumlah sel yang ditekan saat menegakkan ambang batas pada risiko rata-rata ketat relatif terhadap jumlah sel yang ditekan saat menerapkan ambang batas pada risiko penuntut. Dapat dilihat bahwa menggunakan risiko rata-rata ketat menghasilkan hasil yang signifikanfisel yang tidak terlalu ditekan, terutama ketikafigurasi dengan quasi-identiident yang lebih sedikitfiers sedang digunakan. Effefek pada distribusi nilai atribut disajikan dalam Gambar 6(b). Berbeda dengan effdll pada jumlah sel yang ditekan, peningkatan yang diperoleh dalam hal Entropi Non-Seragam meningkat dengan jumlah kuasi-identifier. Ini menyiratkan bahwa kualitas data bisa lebih efisienffmeningkat secara efektif dengan menggunakan model privasi yang kurang ketat ketika harus diasumsikan bahwa musuh memiliki banyak latar belakang pengetahuan.

### 3.4. Analisis eksperimental dari risiko-utilitas trade-off disediakan

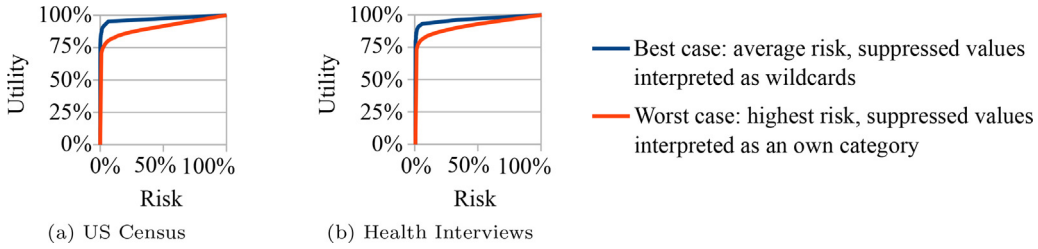
Plugin kami menyediakan spektrum pilihan anonimisasi yang luas, mulai dari parameterisasi yang sangat ketat hingga yang sangat santai. Untuk menganalisis different pilihan secara lebih rinci, kami membangun utilitas risiko

Tabel 1
Perbandingan metode untuk membuat kubus data yang menjaga privasi seperti yang diusulkan oleh Kim et al. [39].

| | Generalisasi global | Generalisasi lokal | Bucketisasi | Penekanan sel |
|---|---|---|---|---|
| kehilangan informasi | 0,41 | 0.13 | Tak dapat diterapkan | 0,10 |
| Kesalahan relatif median untuk kueri titik (%) | 18.3 | 9.79 | 0,02 | 0.00 |
| Kesalahan relatif median untuk kueri rentang (%) | 10.16 | 0,81 | 0,02 | 41.33 |




Gambar 6. Perbandingan hasil yang diperoleh ketika hanya menegakkan ambang batas pada risiko jaksa dengan hasil yang diperoleh menegakkan ambang batas pada risiko rata-rata ketat. Kami melaporkan jumlah sel yang ditekan relatif terhadap jumlah yang diperoleh dengan menggunakan model jaksa untuk kasus di mana setidaknya satu sel ditekan. Kualitas data adalah kuantitasfied menggunakan itu Tidak Seragam model Entropi.




Gambar 7. Batas risiko-utilitas untuk diffmodel risiko yang ada dan diffinterpretasi yang salah dari nilai-nilai yang hilang.

perbatasan, yang plot memvisualisasikan trade-offbahwa metode anonimisasi menyediakan antara perlindungan privasi dan kualitas data [40]. Setiap titik dalam plot ini mewakili kumpulan data yang diubah dariffmenghasilkan pertukaran privasi/utilitas yang optimal, yang berarti bahwa risiko tidak dapat dikurangi lebih lanjut tanpa mengurangi kualitas dan sebaliknya. Gambar 7 menunjukkan hasil metode kami untuk kedua set data menggunakan dua konfigurasi ekstrimfigurasi yang menangani semua quasi-identifier. Dalam skenario kasus terbaik, ambang batas risiko rata-rata $R_{Sebuah}$ telah ditegakkan saat menafsirkan nilai yang hilang sebagai kartu liar. Dalam skenario terburuk, ambang batas pada risiko tertinggi $R_h$ telah ditegakkan sambil memperlakukan nilai-nilai yang hilang sebagai kategori sendiri. Utilitas data diperkirakan dengan jumlah relatif sel yang memiliki tidak telah ditekan.

Seperti yang dapat dilihat, kami tidak dapat mengukur signifikansi apa punfi tidak bisaffperbedaan antara hasil untuk dua set data. Dalam kedua kasus, kami mengamati bahwa kualitas data yang tinggi dapat dipertahankan pada tingkat risiko yang sangat rendah. Batas untuk skenario kasus terbaik hampir optimal. Di sini, kami mengukur area di bawah kurva (AUC, 1 optimal, 0 terburuk) sebesar 0,971 untuk dataset Sensus AS dan 0,966 untuk dataset Health Interviews. Dalam skenario terburuk, kami mengukur AUC masing-masing 0,901 dan 0,912.

### 3.5. Analisis eksperimental effdll dari pemblokiran baris

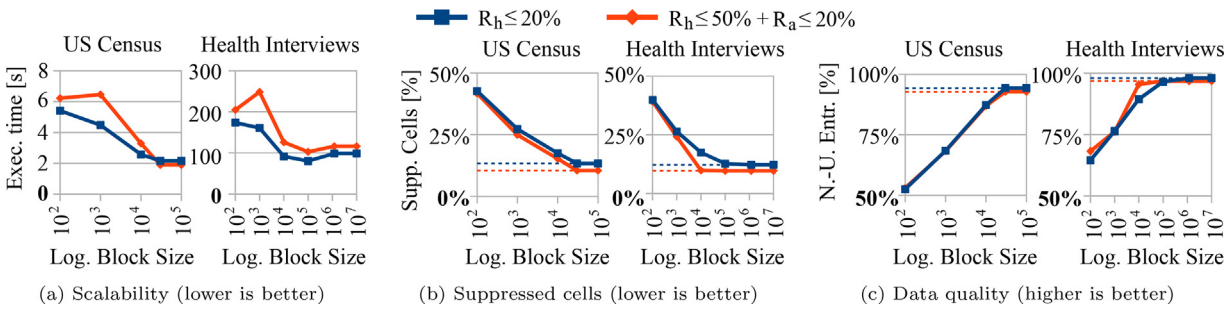
Selanjutnya, kami menyelidiki effdll dari pemblokiran baris pada waktu eksekusi dan kualitas data keluaran. Eksperimen dilakukan dengan sembilan atribut pengidentifikasi semu dan model risiko dan ambang batas yang sama seperti pada eksperimen sebelumnya sambil memvariasikanUkuran blok. Sebelumnya, kami tidak menggunakan pemblokiran baris dan dengan demikian hanya dapat melaporkan waktu yang diperlukan untuk menganonimkan data. Dalam hasil yang disajikan di sini, waktu eksekusi mencakup waktu yang diperlukan untuk membaca data dari disk, menganonimkannya, dan menyimpan hasilnya di disk.

Seperti yang bisa dilihat di Gambar 8(a), waktu eksekusi menurun seiring dengan kekusutan ukuran blok hingga ukuran blok kira-kira $10_5$, dari mana mereka perlahan-lahan meningkat lagi. Peningkatan ini dapat dijelaskan oleh fakta bahwa volume data yang jauh lebih besar diperlukan untuk diproses dalam setiap operasi anonimisasi. Untuk risiko rata-rata ketat dan ukuran blok antara $10_2$ dan $10_3$, kami juga mengamati peningkatan waktu eksekusi. Ini dapat dijelaskan oleh fakta bahwa pengaturan ini signifikanfitidak dapat meningkatkan jumlah pemanggilan dari algoritma anonimisasi yang mendasarinya. Meskipun setiap doa harus menangani sejumlah kecil kerumitan catatan, ko- masalah anonimisasi dengan hormat with ke jumlah kuasi-identifiers tetap konstan. Selain itu, menganonimkan lebih sedikit catatan cenderung lebih mahal secara komputasi, karena solusi yang baik lebih sulit untuk fidan [35]. Mengenai jumlah sel yang ditekan dan effmempengaruhi kualitas data, ketika meningkatkan ukuran blok, kami mengukur penurunan logaritmik ( Gambar 8(b)) dan bertambah (Gambar 8(c)), dengan nilai konvergen menuju garis dasar (bertitik) yang diperoleh tanpa pemblokiran baris. Dengan ukuran blok sekitar $10_4$ (Sensus AS) dan $10_5$ (Wawancara kesehatan) atau lebih besar, effefek pemblokiran baris pada kualitas data keluaran hampir dapat diabaikan dibandingkan dengan anonimisasi tanpa pemblokiran baris. Ini menunjukkan bahwa pemblokiran baris dapat digunakan untuk effsecara efektif menyeimbangkan kualitas data dan waktu eksekusi saat memproses kumpulan data besar.

## 4. Diskusi

### 4.1. Hasil utama dan aplikasi dalam praktik

Pada artikel ini, kami telah menyajikan sebuah plugin yang mendukung anonimisasi data terintegrasi dan identifikasi ulangfianalisis risiko kation selama proses ETL. Implementasi kami didasarkan pada platform PDI, yang digunakan secara luas dalam biomedisfitua. Metode yang disajikan dalam makalah ini juga telah diimplementasikan langsung ke ARX[19]. Metodologi penilaian risiko yang dijelaskan kuat, mudah untukfigure dan memberikan keseimbangan yang baik antara pendekatan sederhana namun ketat seperti k- anonimitas [25] dan banyak lagi flmodel yang fleksibel tetapi kompleks yang memberikan tingkat kualitas data keluaran yang lebih tinggi (misalnya model populasi super [44] atau pendekatan teori permainan [45,46]). Metode transformasi yang diusulkan menghasilkan kumpulan data yang benar yang sangat cocok untuk melakukan analisis statistik umum[21,28,47,34]. Akhirnya, perangkat lunak mengatasi beberapa keterbatasan dari solusi anonimisasi data sebelumnya: data dapat dengan mudah dilindungi dari berbagai ancaman dengan menggabungkan diffoperasi anonimisasi saat ini dalam satu ETL work- flow dan kumpulan data yang sangat besar dapat diproses dengan memanfaatkan model pemrosesan berbasis streaming dari platform yang mendasarinya. Karena kenyataan bahwa pendekatan kami dapat digunakan untuk memproses data yang telah dipartisi menjadi subset independen (lihat Bagian 2.3 dan Bagian C dari pelengkapfile) itu juga dapat digunakan untuk menambahkan data secara bertahap ke




Gambar 8. Plot semi-log memvisualisasikan hasil eksperimen pemblokiran baris. Kami melaporkan waktu eksekusi rata-rata, jumlah sel yang ditekan, dan kualitas data seperti yang dilaporkan oleh model Entropi Non-Seragam. Garis putus-putus mewakili nilai dasar yang diperoleh tanpa pemblokiran baris.

database yang ada tanpa melanggar jaminan privasi yang diberikan.

Perangkat lunak yang dijelaskan dalam artikel ini telah digunakan di berbagai proyek. Misalnya, digunakan untuk menganonimkan data demografis untuk gudang data penelitian di Departemen Penyakit Kardiovaskular Pusat Jantung Jerman Munich. Gudang tersebut mengintegrasikan data fenotipik dan genotipik lebih dari 70.000 pasien dengan penyakit arteri koroner untuk mendukung visualisasi data, penemuan kohort, dan pembuatan hipotesis. Kami juga sering menggunakan metodologi yang dijelaskan di sini saat melindungi ekstrak data sebelum membagikannya dengan mitra eksternal, misalnya dalam konteks pendaftar penelitian untuk gangguan mitokondria[48] dan untuk penyakit neurodegeneratif [49]. Akhirnya, metode yang dijelaskan juga telah digunakan melalui ARX oleh kelompok penelitian lain, misalnya untuk membuat kumpulan data terbuka untuk studi perilaku belajar.[50] dan untuk menganonimkan data dari program skrining kanker [51].

### 4.2. Perbandingan konseptual dengan pekerjaan sebelumnya

Pada tingkat konseptual, pekerjaan sebelumnya dapat ditemukan di banyak bidang, termasuk anonimisasi data, pembuatan data sintetis, dan penyamaran data. Kami telah membahas lingkungan terkait untuk menerapkan proses ETL dan solusi anonimisasi data sumber terbuka lainnya di bagian sebelumnya. Perangkat lunak lain yang layak disebut adalah Privacy Analytics Eclipse[52], yang merupakan platform anonimisasi data komersial yang dibangun di atas Apache Spark [53]. Sementara perangkat lunak mengimplementasikan metode formal yang sangat mirip dengan yang diterapkan oleh plugin kami, sedikit yang telah dipublikasikan tentang metodologi yang tepat dan implementasinya.

Di sisa bagian ini, kami fokus pada solusi lebih lanjut yang mengintegrasikan fitur perlindungan data ke dalam proses ETL. Penyamaran data adalah teknik yang juga telah diintegrasikan ke dalam platform ETL. Metode dari inifilapangan tidak didasarkan pada penilaian risiko formal dan anonimisasi data, tetapi mereka menerapkan proses transformasi berbasis aturan sederhana, misalnya untuk penghapusan data. Mereka biasanya digunakan untuk membuat data untuk pengembangan perangkat lunak dan tujuan pengujian. Contoh implementasi yang relevan termasuk Informatica'sPenyamaran Data [54], Privasi Data IBM InfoSphere Optim [55], Paket Penyembunyian dan Subsetting Data Oracle [56], ProxySQL [57], dan Komponen Data Masking Hush Hush [58]. Juga, TOS dan PDI keduanya offer modul yang menyediakan fungsionalitas penyembunyian data dasar.

Pembuatan data sintetis juga didukung oleh solusi masking yang disajikan dalam paragraf sebelumnya. Sebagian besar implementasi agak sederhana, tetapi ada juga pendekatan yang canggih, seperti algoritma yang didukung oleh sdcMicro[18] yang mampu mempertahankan sifat statistik uni dan multivariat dari data masukan. Pembuatan data acak juga didukung oleh plugin untuk TOS dan PDI. Bijoux adalah contoh terkenal lainnya[59]. Namun, plugin pembuatan data untuk proses ETL biasanya terlalu sederhana untuk berguna lebih dari sekadar pembuatan data uji.

## 5. Kesimpulan dan pekerjaan masa depan

Dalam artikel ini, kami telah menjelaskan plugin untuk platform ETL umum yang mendukung fungsi anonimisasi dan penilaian risiko yang kuat. Perangkat lunak ini tersedia di bawah lisensi open source non-restriktif. Metode kami dapat diintegrasikan ke dalam pekerjaan ETL yang adaflows, dan mendukung solusi pergudangan khas untuk data biomedis, seperti i2b2 dan transSMART. Bahkan dalam kasus di mana tidak mungkin untuk menandatanganifitidak dapat mengurangi risiko tanpa dampak yang cukup besar pada utilitas data, perangkat lunak kami dapat digunakan untuk melakukan identifikasi ulang kuantitatiffi penilaian risiko kation untuk mendokumentasikan ancaman privasi. Ini adalah aspek penting dari undang-undang privasi modern, seperti Peraturan Perlindungan Data Umum Eropa[16].

Metode dan implementasi yang disajikan dalam artikel ini sangat cocok untuk melindungi data yang jarang dikumpulkan (misalnya demografi) atau yang tetap agak stabil dari waktu ke waktu (misalnya diagnosis atau nilai lab yang menarik untuk suatufi belajar) [14,27]. Jika data longitudinal atau sering berubah perlu dilindungi dari serangan linkage, spesifikasikanfic langkah-langkah harus diterapkan yang dapat mengatasi dimensi dan perubahan data yang lebih tinggi [60]. Sementara kami berencana untuk memperluas perangkat lunak kami untuk mencakup kasus penggunaan seperti itu dalam pekerjaan di masa depan, kami juga menekankan bahwa data tersebut sering menimbulkan risiko yang jauh lebih kecil, karena tidak stabil,FFIkultus untuk ditiru dan karena itu kecil kemungkinannya bahwa pengetahuan latar belakang yang memadai tersedia untuk musuh [14,27]. Area tambahan pekerjaan di masa depan adalah dukungan yang ditingkatkan untuk menambahkan data baru secara bertahap. Meskipun ini sudah didukung oleh versi plugin kami saat ini, kami berencana untuk menambahkan fungsionalitas untuk mempertimbangkan data yang sudah ada dalam database saat mengukur dan mengurangi risiko selama proses memuat data baru. Ini dapat membantu untuk lebih mengurangi jumlah penekanan yang diperlukan.

Penindasan sel memungkinkan anonimisasi kumpulan data dengan koni minimalfigurasifforts, tetapi metode transformasi lebih lanjut juga dapat berguna dalam skenario tertentu. Generalisasi data dan mikroagregasi adalah dua teknik spesifikfibunga c. Kami berencana untuk menambahkan dukungan di versi plugin yang akan datang. Namun, karena metode ini mungkin berdampak pada sifat skema data (misalnya perubahan tipe data dan skala pengukuran) mengintegrasikannya dengan lingkungan pemrosesan solusi ETL merupakan tantangan. Pendekatan anonimisasi alternatif untuk penekanan sel adalah pertukaran sel (atau pertukaran data)[61] yang pada dasarnya bekerja dengan menukar nilai atribut antar record. Analog dengan pekerjaan kami, itu mempertahankan sifat skematis data. Berbeda dengan pendekatan kami, pertukaran data tidak menghapus nilai atribut dan karenanya mempertahankan agregat statistik seperti jumlah nilai atribut. Namun, tidak seperti penekanan sel, pertukaran data secara inheren mengganggu. Oleh karena itu, itu tidak memenuhi kebenaran, yang merupakan persyaratan penting dalam konteks kita (lih. Bagian 2.2). Selain itu, pertukaran data biasanya diimplementasikan berdasarkan model risiko sederhana, yang tidakffer tingkat perlindungan yang jauh lebih rendah daripada metode yang digunakan dalam pekerjaan kami. Arah potensial untuk pekerjaan di masa depan adalah menyelidiki bagaimana pertukaran data dapat diintegrasikan ke dalam kerangka anonimisasi yang diusulkan, termasuk yang kuat

model perlindungan yang digunakan, dan untuk memeriksa potensi peningkatan utilitas data yang dihasilkan. Salah satu pendekatan yang mungkin untuk ini adalah denganfi pertama-tama lakukan anonimisasi menggunakan penekanan sel, termasuk penilaian risiko seperti yang dijelaskan dalam artikel ini. Langkah ini kemudian dapat diikuti oleh langkah pascapemrosesan di mana nilai asli dari sel yang ditekan ditukar dan kemudian dimasukkan kembali ke dalam kumpulan data keluaran.

Sementara plugin kami mendukung salah satu lingkungan yang paling banyak diadopsi untuk menerapkan proses ETL, TOS juga sering digunakan dalam proyek pergudangan data biomedis. Kami sudah mulai mem-port plugin kami ke platform ini tetapi, karena perbedaanffperbedaan antara lingkungan pengembangan dan konsep untuk mengelola data dan kontrol
flows, integrasi lengkap akan membutuhkan lebih banyak pekerjaan.

## Poin ringkasan

Apa yang sudah diketahui tentang topik itu?

- Anonimisasi penting dalam penelitian biomedis, terutama ketika data dikumpulkan atau digunakan kembali untuk tujuan sekunder.
- Alat ETL (Extract-Transform-Load) umum untuk mengintegrasikan data ke gudang klinis dan translasi tidak mendukung anonimisasi. Selain itu, alat anonimisasi umum tidak dapat dengan mudah terintegrasi ke dalam pekerjaan ETLfladuh.
- Alat anonimisasi dapat di canFFIkultus untuk konfigure dan mereka punya masalah skalabilitas saat memproses kumpulan data yang sangat besar.

Apa yang telah ditambahkan penelitian ini ke dalam tubuh pengetahuan?

- Metodologi anonimisasi tingkat ahli dapat diintegrasikan sebagai plugin tuitif ke dalam platform ETL.
- Dengan plugin ini, data dapat dilindungi dari berbagai ancaman dalam satu pekerjaan ETLfladuh.
- Kumpulan data yang sangat besar dapat dianonimkan eFFIsecara efisien dengan memanfaatkan model pemrosesan platform ETL berbasis streaming.
- Utilitas dan kompatibilitas data tinggi dengan database yang ada dan platform dapat dicapai dengan menggunakan metode transformasi yang menjaga kebenaran data dan sifat skematisnya.

;1;

## Pengarang' kontribusi

FP, HS, dan RB mengembangkan algoritma. HS dan FP merancang dan mengimplementasikan plugin. HS, JE, RB dan FP merancang, mengimplementasikan dan melakukan eksperimen. RB, HS dan FP mengembangkan bukti formal kebenaran pendekatan. HS, RB, JE, KK dan FP membahas konsepsi dan desain karya serta naskah di semua tahapan. Semua penulis telah berkontribusi pada naskah. Semua penulis telah membaca dan menyetujuifinaskah akhir.

## Kepentingan bersaing

Para penulis menyatakan bahwa mereka tidak memiliki kepentingan yang bersaing.

## Ucapan Terima Kasih

Pekerjaan itu, sebagian, didanai oleh Kementerian Pendidikan dan Penelitian Federal Jerman (BMBF) di dalam “Skema Pendanaan Informatika Medis” di bawah nomor referensi 01ZZ1804A (DIFUTURE).

## Lampiran A. Data tambahan

Data tambahan yang terkait dengan artikel ini dapat ditemukan, di: versi online, di https://doi.org/10.1016/j.ijmedinf.2019.03.006.

## Referensi

[1] SO Tanggul, AA Philippakis, JR De Argila, DN Paltoo, ES Luetkemeier, BM Knoppers, AJ Brookes, JD Spalding, M. Thompson, M. Roos, dkk., Kode izin: menjunjung tinggi kondisi penggunaan data standar, PLoS Genet. 12 (1) (2016) e1005772,https://doi.org/10.1371/journal.pgen.1005772.

[2] S. Schneeweiss, Belajar dari data perawatan kesehatan besar, N. Engl. J. Med. 370 (23) (2014) 2161-2163, https://doi.org/10.1056/NEJMp1401111.

[3] AJ McMurry, SN Murphy, D. MacFadden, G. Weber, WW Simons, J. Orechia, J. Bickel, N. Wattanasin, C. Gilbert, P. Trevvett, et al., SHRINE: memungkinkan studi penyakit multi-lokasi yang dapat diskalakan secara nasional, PLoS One 8 (3) (2013) e55811 https://doi.org/10.1371/journal.pone.0055811.

[4] K. Shameer, MA Badgeley, R. Miotto, BS Glicksberg, JW Morgan, JT Dudley, Bioinformatika translasi di era aliran data biomedis, perawatan kesehatan, dan kebugaran real-time, Singkat. Informasi biologis. 18 (1) (2017) 105-124, https://doi.org/10. 1093/bib/bbv118.

[5] I. Danciu, JD Cowan, M. Basford, X. Wang, A. Saip, S. Osgood, J. Shirey-Rice, J. Kirby, PA Harris, Penggunaan sekunder data klinis: pendekatan Vanderbilt, J. Biomed. Memberitahu. 52 (2014) 28-35, https://doi.org/10.1016/j.jbi.2014.02.003.

[6] A.-S. Jannot, E. Zapletal, P. Avilach, M.-F. Mamzer, A. Burgun, P. Degoulet, Gudang Data Klinis Rumah Sakit Universitas Georges Pompidou: pengalaman tindak lanjut 8 tahun, Int. J. Med. Memberitahu. 102 (2017) 21-28, https://doi.org/10.1016/j. ijmedinf.2017.02.006.

[7] SN Murphy, G. Weber, M. Mendis, V. Gainer, HC Chueh, S. Churchill, I. Kohane, Melayani perusahaan dan seterusnya dengan informatika untuk mengintegrasikan biologi dan samping tempat tidur (i2b2), J. Am. Med. Memberitahu. Asosiasi 17 (2) (2010) 124-130, https://doi.org/10.1136/jamia.2009.000893.

[8] E. Scheufele, D. Aronzon, R. Coopersmith, MT McDuFFIe, M. Kapoor, CA Uhrich, JE Avitabile, J. Liu, D. Housman, MB Palchuk, tranSMART: manajemen pengetahuan sumber terbuka dan platform analisis data konten tinggi, AMIA Jt. Terjemahan KTT Sci. Prok. (2014) 96-101.

[9] W. Inmon, Membangun Gudang Data, John Wiley & Sons, 2005.

[10] MJ Denney, DM Long, MG Armistead, JL Anderson, BN Conway, Memvalidasi proses ekstrak, transformasi, pemuatan yang digunakan untuk mengisi database penelitian klinis besar, Int. J. Med. Memberitahu. 94 (2016) 271-274.

[11] M. Casters, R. Bouman, J. Van Dongen, Pentaho Kettle Solutions: Membangun Solusi ETL Open Source dengan Integrasi Data Pentaho, John Wiley & Sons, 2010.

[12] J. Bowen, Memulai Talend Open Studio untuk Integrasi Data, Packt Publishing Ltd, 2012.

[13] C. Bauer, T. Ganslandt, B. Baum, J. Christoph, I. Engel, M. Löbe, S. Mate, S. Stäubert, J. Drepper, H.-U. Prokosch, U. Sax, Perangkat Penyimpanan Data Terintegrasi (IDRT). Serangkaian program untuk memfasilitasi analisis kesehatan pada data medis yang heterogen, Metode Inf. Med. 55 (2) (2016) 125-153, https://doi.org/10.3414/ME1501-0082.

[14] BA Malin, D. Karp, RH Scheuermann, Pendekatan teknis dan kebijakan untuk menyeimbangkan privasi pasien dan berbagi data dalam penelitian klinis dan translasi, J. Investig. Med. 58 (1) (2010) 11-18, https://doi.org/10.2310/JIM. 0b013e3181c9b2ea.

[15] Departemen Kesehatan dan Layanan Kemanusiaan AS, Standar untuk privasi identitas individufiinformasi kesehatan dapat, Aturan Akhir. 45 CFR, Bagian 160-164, Daftar Federal 67 (157) (2002) 53182-53273.

[16] Peraturan (UE) 2016/679 dari Eur. Parlemen dan Dewan 27 April 2016 tentang perlindungan orang perseorangan sehubungan dengan pemrosesan data pribadi dan tentang pergerakan bebas data tersebut, dan mencabut arahan 95/46/EC (Peraturan Perlindungan Data Umum), O. J.Eur. Serikat (Mei 2016) L119/59.

[17] F. Kohlmayer, F. Prasser, KA Kuhn, Biaya kualitas: menerapkan generalisasi dan penekanan untuk menganonimkan data biomedis dengan kehilangan informasi minimal, J. Biomed. Memberitahu. 58 (2015) 37-48, https://doi.org/10.1016/j.jbi.2015.09. 007.

[18] M. Templ, A. Kowarik, B. Meindl, Kontrol pengungkapan statistik untuk microdata menggunakan paket R sdcMicro, J. Stat. lunak 67 (1) (2015) 1-36, https://doi.org/10. 18637/jss.v067.i04.

[19] F. Prasser, F. Kohlmayer, Menerapkan kontrol pengungkapan statistik ke dalam praktik: alat anonimisasi data ARX, Buku Pegangan Privasi Data Medis, Springer, 2015, hlm. 111-148.

[20] K. El Emam, L. Arbuckle, Anonimisasi Data Kesehatan: Studi Kasus dan Metode untuk Anda Memulai, edisi pertama, O'Reilly, 2013.

[21] K. El Emam, BA Malin, Lampiran B: Konsep dan metode untuk mengidentifikasi data uji klinis, dalam: Komite Strategi untuk Berbagi Data Uji Coba Klinis yang Bertanggung Jawab, Dewan Kebijakan Ilmu Kesehatan, Institute of Medicine (Eds.), Berbagi Data Uji Klinis: Memaksimalkan Manfaatfits, Meminimalkan Risiko, National Academies Press (AS), Washington (DC), 2015, hlm-290.

[22] Badan Uni Eropa untuk Keamanan Jaringan dan Informasi (ENISA), Privasi dan Perlindungan Data oleh Desain - dari kebijakan ke rekayasa (2014), 1-79.

[23] Badan Obat Eropa (EMA), EMA/90915/2016 - Panduan eksternal tentang penerapan kebijakan Badan Obat Eropa tentang publikasi data klinis untuk produk obat untuk penggunaan manusia (2016), 1-99.

[24] B. Fung, K. Wang, R. Chen, PS Yu, Penerbitan data pelestarian privasi: survei perkembangan terakhir, ACM Comput. bertahan (CSUR) 42 (4) (2010) 14.

[25] L. Sweeney, k-anonymity: model untuk melindungi privasi, Int. J. Ketidakpastian Ketidakjelasan Sistem Berbasis Pengetahuan. 10 (05) (2002) 557-570.

[26] F. Prasser, F. Kohlmayer, KA Kuhn, dkk., Pentingnya konteks: de-identi berbasis risikofi kation data biomedis, Metode Inf. Med. 55 (4) (2016) 347-355, https://doi.org/10.3414/ME16-01-0012.

[27] B. Malin, G. Loukides, K. Benitez, EW Clayton, Identifikemampuan dalam biobank: model,

langkah-langkah, dan strategi mitigasi, Hum. gen. 130 (3) (2011) 383.
[28] K. El Emam, Panduan untuk De-Identifikation Informasi Kesehatan Pribadi, CRC Press, 2013 .
[29] DC Barth-Jones, The 'Identifikasi Ulangfikation' Informasi Medis Gubernur William Weld: pemeriksaan ulang kritis terhadap identitas data kesehatanfirisiko kation dan perlindungan privasi, dulu dan sekarang, Tersedia dari SSRN: http://ssrn.com/abstract=2076397. Diakses 5 Januari 2018 (2012). doi:10.2139/ssrn.2076397.
[30] FK Dankar, K. El Emam, Berlatih diffprivasi penting dalam perawatan kesehatan: ulasan, Trans. Privasi Data 6 (1) (2013) 35-67.
[31] C. Dwork, Diffprivasi penting: survei hasil, Konferensi Internasional tentang Teori dan Aplikasi Model Komputasi Springer (2008) 1-19.
[32] J. Domingo-Ferrer, JM Mateo-Sanz, Agregasi mikro berorientasi data praktis untuk kontrol pengungkapan statistik, IEEE Trans. tahu. Data Eng. 14 (1) (2002) 189-201.
[33] X. Xiao, Y. Tao, Anatomi: sederhana dan effpelestarian privasi yang efektif, Prosiding Konferensi Internasional ke-32 tentang Basis Data Sangat Besar, VLDB Endowment, 2006, hlm. 139-150.
[34] L. Ohno-Machado, S. Vinterbo, S. Dreiseitl, Effefek anonimisasi data dengan penekanan sel pada statistik deskriptif dan kinerja pemodelan prediktif, J. Am. Med. Memberitahu. Asosiasi 9 (Tambahan_6) (2002) S115-S119.
[35] F. Prasser, F. Kohlmayer, KA Kuhn, EFFIilmiah dan effstrategi pemangkasan yang efektif untuk de-identifikasi data kesehatanfikation, BMC Med. Memberitahu. keputusan Mak. 16 (1) (2016) 49, https://doi.org/10.1186/s12911-016-0287-2.
[36] arx-deidentifier/arx-pdi-plugins, Plugin untuk platform Integrasi Data Pentaho. tersedia dari https://github.com/arx-deidentifier/arx-pdi-plugin. Diakses pada 23 Maret 2018.
[37] arx-deidentifier/cell-suppression-benchmark, Tolok ukur metode penekanan sel di ARX. tersedia darihttps://github.com/arx-deidentifier/penindasan sel-benchmark. Diakses pada 23 Maret 2018.
[38] M. Ciglic, J. Eder, C. Koncilia, k-Anonimitas mikrodata dengan nilai NULL, Int. Kon. Database Eks. Sis. aplikasi Musim Semi (2014) 328-342, https://doi.org/10.1007/978-3-319-10073-9_27.
[39] S. Kim, H. Lee, YD Chung, kubus data pelestarian privasi untuk rekam medis elektronik: evaluasi eksperimental, Int. J. Med. Memberitahu. 97 (2017) 33-42, https://doi.org/10.1016/j.ijmedinf.2016.09.008.
[40] LH Cox, AF Karr, SK Kinney, Paradigma utilitas risiko untuk pembatasan pengungkapan statistik: cara berpikir, tetapi bukan cara bertindak, Int. Stat. Wahyu 79 (2) (2011) 160-183, https://doi.org/10.1111/j.17515823.2011.00140.x.
[41] F. Prasser, F. Kohlmayer, KA Kuhn, EFFIilmiah dan effstrategi pemangkasan yang efektif untuk de-identifikasi data kesehatanfikation, BMC Med. Memberitahu. keputusan Mak. 16 (1) (2016) 49, https://doi.org/10.1186/s12911-016-0287-2.
[42] F. Prasser, R. Bild, J. Eicher, H. Spengler, F. Kohlmayer, KA Kuhn, Lightning: anonimisasi berbasis utilitas dari data dimensi tinggi, Trans. Privasi Data 9 (2) (2016) 161-185.
[43] A. De Waal, L. Willenborg, Kehilangan informasi melalui pengodean ulang global dan penekanan lokal, Belanda O. Stat. 14 (1999) 17-20.
[44] FK Dankar, K. El Emam, A. Neisa, T. Roffey, Memperkirakan re-identifirisiko kation set data klinis, BMC Med. Memberitahu. keputusan Mak. 12 (1) (2012) 66,https://doi.org/10.1186/1472-6947-12-66.
[45] Z. Wan, Y. Vorobeychik, W. Xia, EW Clayton, M. Kantarcioglu, R. Ganta, R. Heatherly, BA Malin, Kerangka teori permainan untuk menganalisis identifikasi ulangfi risiko kation, PLoS One 10 (3) (2015) e0120592, https://doi.org/10.1371/journal.pon.0120592.
[46] F. Prasser, J. Gaupp, Z. Wan, W. Xia, Y. Vorobeychik, M. Kantarcioglu, KA Kuhn, BA Malin, Alat open source untuk de-identi data kesehatan teoretis gamefikation, AMIA Annu. Sim. Prok. (2017).
[47] K. El Emam, FK Dankar, R. Issa, E. Jonker, D. Amyot, E. Cogo, J.-P. Corriveau, M. Walker, S. Chowdhury, R. Vaillancourt, dkk., Metode k-anonimitas optimal global untuk de-identifikasi data kesehatan, J. Am. Med. Memberitahu. Asosiasi 16 (5) (2009) 670-682, https://doi.org/10.1197/jamia.M3144.
[48] B. Büchner, C. Gallenmüller, R. Lautenschläger, K. Kuhn, I. Wittig, L. Schöls, D. Rapaport, D. Seelow, P. Freisinger, H. Prokisch, dkk., Jaringan Jerman untuk Gangguan Mitokondria (mitoNET), Med. gen. 24 (3) (2012) 193-199, https://doi.org/10.1007/s11825-012-0338-8.
[49] B. Kalman, B. Büchner, F. Kohlmayer, KA Kuhn, R. Lautenschlaeger, T. Klopstock, T. Kmiec, Registri internasional untuk neurodegenerasi dengan akumulasi besi otak, Orphanet J. Rare Dis. 7 (2012) 66,https://doi.org/10.1186/1750-11727-66.
[50] J. Kuzilek, M. Hlosta, Z. Zdrahal, kumpulan data Analisis Pembelajaran Universitas Terbuka, Sci. Data 4 (2017) 170171,https://doi.org/10.1038/sdata.2017.171.
[51] G. Ursin, S. Sen, J.-M. Mottu, M. Nygård, Melindungi privasi dalam kumpulan data besar—fipertama kita menilai risikonya; lalu kami mengaburkan datanya, Cancer Epidemiol. Biomarker Sebelumnya 26 (8) (2017) 1219-1224, https://doi.org/10.1158/1055-9965.EPI-17-0172.
[52] Analisis Privasi, Inc., Eclipse Analisis Privasi. tersedia darihttps://privacyanalytics.com/software/privacy-analytics-Eclipse/. Diakses pada 5 Januari 2018.
[53] Apache Spark. tersedia darihttps://spark.apache.org/. Diakses 12 Januari 2018.
[54] Perusahaan Informatica, Penyembunyian Data. tersedia darihttps://www.informatika. com/gb/products/data-security/data-masking.html. Diakses pada 5 Januari 2018.
[55] IBM Corporation, Privasi Data IBM InfoSphere Optim. tersedia darihttps://www.ibm.com/ms-en/marketplace/infosphere-optim-data-privacy/details# product-header-top. Diakses pada 5 Januari 2018.
[56] Oracle Corporation, Oracle Data Masking and Subsetting Pack. tersedia dari http://www.Oracle.com/technetwork/database/options/data-masking-subsetting/Overview/ds-security-dms-2245926.pdf. Diakses 12 Januari 2018 (2016).
[57] R. Cannaò, ProxySQL. tersedia darihttp://proxysql.com. Diakses 5 Januari 2018 (2018).
[58] Hush, Teknologi dan Layanan Informasi Hush, Komponen Penyembunyian Data. tersedia darihttp://mask-me.net/. Diakses 5 Januari 2018 (2017).
[59] V. Theodorou, P. Jovanovic, A. Abelló, E. Nakuçi, Generator data untuk mengevaluasi kualitas proses ETL, Inform. Sis. 63 (Tambahan C) (2017) 80-100, https://doi.org/10. 1016/j.is.2016.04.005.
[60] M. Terrovitis, N. Mamoulis, P. Kalnis, anonimisasi privasi-pemeliharaan dari data yang ditetapkan, Proceedings of the VLDB Endowment 1 (1) (2008) 115-125.
[61] SE Fienberg, J. McIntyre, Data swapping: variasi pada tema oleh Dalenius dan Reiss, J. O. Stat. 21 (2) (2005) 309.