Ong dkk. Informatika Medis BMC dan Pengambilan Keputusan (2017) 17:134
DOI 10.1186/s12911-017-0532-3

BMC Medical Informatics and Decision Making

KEMAJUAN TEKNIS Akses terbuka

Dynamic-ETL: pendekatan hibrida untuk ekstraksi, transformasi, dan pemuatan data kesehatan data

Toan C. Ong[1*], Michael G. Kahn[1,4], Bethany M. Kwan[2], Traci Yamashita[3], Elias Brandt[5], Patrick Hosokawa[2], Chris Uhrich[6] dan Lisa M. Schilling[3]

Abstrak

Latar Belakang: Catatan kesehatan elektronik (EHR) berisi data klinis terperinci yang disimpan dalam format berpemilik dengan kode dan struktur non-standar. Berpartisipasi dalam jaringan penelitian klinis multi-situs membutuhkan data EHR untuk direstrukturisasi dan diubah menjadi format umum dan terminologi standar, dan secara optimal terkait dengan sumber data lain. Keahlian dan solusi terukur yang diperlukan untuk mengubah data agar sesuai dengan persyaratan jaringan berada di luar cakupan banyak organisasi perawatan kesehatan dan ada kebutuhan akan alat praktis yang menurunkan hambatan kontribusi data ke jaringan penelitian klinis.

Metode: Kami merancang dan menerapkan pendekatan transformasi dan pemuatan data kesehatan, yang kami sebut sebagai Dynamic ETL (Ekstraksi, Transformasi, dan Pemuatan) (D-ETL), yang mengotomatiskan bagian dari proses melalui penggunaan kode yang dapat diskalakan, dapat digunakan kembali, dan dapat disesuaikan, sambil mempertahankan aspek manual dari proses yang membutuhkan pengetahuan tentang sintaks pengkodean yang kompleks. Pendekatan ini memberikan fleksibilitas yang diperlukan untuk ETL data heterogen, variasi dalam keahlian semantik, dan transparansi logika transformasi yang penting untuk menerapkan konvensi ETL di seluruh jaringan berbagi penelitian klinis. Alur kerja pemrosesan diarahkan oleh pedoman spesifikasi ETL, yang dikembangkan oleh desainer ETL dengan pengetahuan luas tentang struktur dan semantik data kesehatan (yaitu,"pakar domain data kesehatan") dan menargetkan model data umum.

Hasil: D-ETL diimplementasikan untuk melakukan operasi ETL yang memuat data dari berbagai sumber dengan struktur skema database yang berbeda ke dalam model data umum Observational Medical Outcome Partnership (OMOP). Hasil menunjukkan bahwa metode komposisi aturan ETL dan mesin D-ETL menawarkan solusi terukur untuk transformasi data kesehatan melalui pembuatan kueri otomatis untuk menyelaraskan kumpulan data sumber.

Kesimpulan: D-ETL mendukung proses yang fleksibel dan transparan untuk mengubah dan memuat data kesehatan menjadi model data target. Pendekatan ini menawarkan solusi yang menurunkan hambatan teknis yang mencegah mitra data untuk berpartisipasi dalam jaringan data penelitian, dan oleh karena itu, mendorong kemajuan penelitian efektivitas komparatif menggunakan data kesehatan elektronik sekunder.

Kata kunci: Catatan kesehatan elektronik, Ekstraksi, Transformasi dan pemuatan, Jaringan penelitian terdistribusi, Harmonisasi data, ETL berbasis aturan

* Korespondensi: Toan.Ong@ucdenver.edu
[1]Departemen Pediatri, Kampus Medis Universitas Colorado Anschutz, Fakultas Kedokteran, Gedung AO1 Room L15-1414, 12631 East 17th Avenue, Mail Stop F563, Aurora, CO 80045, USA
Daftar lengkap informasi penulis tersedia di akhir artikel

Latar Belakang

Data klinis - seperti dari catatan kesehatan elektronik (EHR) - telah menjadi sumber data utama (yaitu, sebagai data sekunder) untuk penelitian efektivitas komparatif (CER) [1, 2]. Komunitas penelitian klinis telah lama membayangkan menggunakan data yang dihasilkan selama perawatan klinis rutin untuk mengeksplorasi pertanyaan perawatan kesehatan yang bermakna dan masalah kebijakan kesehatan yang tidak dapat ditangani oleh uji klinis acak tradisional [3-7]. Perkembangan terbaru dalam CER observasional, metode penelitian hasil yang berpusat pada pasien (PCOR), dan teknik analitik telah meningkatkan kemampuan untuk menyimpulkan asosiasi yang valid dari studi observasional non-acak [8-14]. Tujuan saat ini dari beberapa inisiatif data kesehatan utama AS adalah untuk membuat jaringan data besar yang mendukung CER dengan mengintegrasikan data EHR dari berbagai sumber (yaitu, beberapa EHR dari berbagai organisasi perawatan kesehatan) dan memperkaya data ini dengan data klaim [6, 13-19]. Untuk menyelaraskan data dari berbagai sumber, jaringan data kesehatan mengubah data dari sistem EHR sumber menjadi model data umum (CDM), seperti yang dimiliki oleh Observational Medical Outcomes Partnership (OMOP), Informatika untuk Mengintegrasikan Biologi dan Bedside (i2b2), MiniSentinel (MS) dan Jaringan Penelitian Hasil Berpusat pada Pasien (PCORNet) [16, 20-24].

Proses harmonisasi data diketahui menghabiskan sumber daya yang signifikan, dan banyak pekerjaan sebelumnya telah dilakukan untuk menyederhanakan pemetaan data, mempersingkat waktu kueri data, dan meningkatkan kualitas data [25-28]. Tantangan teknis umum dari proses ETL adalah kompatibilitas sumber dan data target, skalabilitas proses ETL, dan kualitas data sumber [29-33]. Tantangan kompatibilitas terjadi karena sistem EHR lokal sering kali memiliki model data, kosakata, istilah untuk elemen data, dan tingkat granularitas data yang berbeda. Masalah ketidakcocokan dapat menyebabkan hilangnya informasi karena ketidakmampuan model data target untuk menerjemahkan dan menyimpan sintaks dan semantik dari data sumber secara akurat [34]. Skalabilitas merupakan tantangan karena volume data kesehatan, kebutuhan untuk penyegaran data yang sering, perubahan operasional dalam sistem data sumber, dan revisi berkelanjutan untuk menargetkan definisi dan cakupan skema. Akhirnya, memastikan kualitas data sebagai hasil dari proses ETL merupakan tantangan karena berbagai kualitas sumber data EHR yang bergantung pada organisasi sumber.' Implementasi EHR dan interaksi pengguna akhir dengan sistem [35]. Tantangan transformasi data lainnya melibatkan penyediaan solusi untukbertentangan dan duplikat catatan. bertentangan

catatan didefinisikan sebagai dua atau lebih catatan tentang objek yang sama (misalnya pasien, kunjungan) yang memiliki identifikasi yang sama (misalnya nomor pertemuan yang sama) tetapi menyatakan nilai yang berbeda untuk fakta atau pengamatan yang diberikan. Di samping itu, duplikat catatan mengacu pada dua catatan yang memiliki nilai identik di semua kolom kecuali catatan kunci utama pengenal. Catatan yang saling bertentangan dan duplikat adalah masalah data umum yang dapat secara signifikan mempengaruhi efisiensi proses ETL dan kualitas data keluaran. Pendekatan saat ini untuk transformasi data seringkali tidak fleksibel atau terukur untuk inisiatif besar dengan banyak sumber data yang heterogen dan hubungan yang sangat spesifik antara elemen data dalam model data sumber dan target [30, 36, 37].

Proses ETL (Extraction-Transformation-Load) adalah serangkaian operasi yang memungkinkan data sumber diharmonisasikan secara sintaksis dan semantik dengan struktur dan terminologi CDM target [38]. Proses ETL untuk mendukung harmonisasi data biasanya terdiri dari dua fase berurutan, yang masing-masing dilakukan oleh personel yang terampil dengan keahlian yang berbeda. Pada fase 1, ahli materi pelajaran dalam data sumber (misalnya EHR, data klaim) mengidentifikasi elemen data yang sesuai yang diperlukan untuk mengisi database target untuk ekstraksi dan menentukan pemetaan antara data sumber dan elemen data target. Langkah ini membutuhkan pengetahuan tentang struktur dan semantik dari data sumber dan target, seperti keahlian dalam implementasi dan penggunaan EHR lokal, dan terminologi lokal. Pada fase 2, pemrogram database menerapkan metode transformasi data dan pemetaan skema untuk memuat data ke dalam skema yang diselaraskan. Transformasi adalah proses data yang kompleks“pembersihan”

(misalnya, de-duplikasi data, resolusi konflik) dan standarisasi (misalnya pemetaan terminologi lokal) agar sesuai dengan format dan kode skema target sehingga dapat dimuat ke dalam basis data spesifik CDM target. Fase ini membutuhkan pemrograman secara manual menggunakan bahasa pemrograman database seperti Structured Query Language (SQL). Dalam banyak kasus, langkah-langkah ini diulang sampai data yang diubah diterima sebagai lengkap dan benar. Kedua fase ini (pemetaan skema; pemrograman database) harus dilakukan secara terpisah untuk setiap sumber data, dan jarang ada satu orang yang memiliki keahlian data sumber dan keterampilan pemrograman database untuk melakukan tugas di kedua fase bahkan untuk satu sumber data, dan terutama tidak untuk beberapa sumber data.

Proses ETL dapat didukung oleh alat integrasi data dengan antarmuka pengguna grafis (GUI), seperti Talend[1] dan Pentaho,[2] yang membantu mengurangi beban manual dari proses desain ETL. Namun, alat berbasis GUI seringkali tidak cukup fleksibel untuk mengatasi persyaratan rumit dari operasi transformasi seperti konvensi khusus untuk melakukan de-duplikasi data atau untuk melakukan beban data tambahan. Juga, alat berbasis GUI sering kekurangan transparansi perintah SQL yang mendasari melakukan transformasi sehingga sulit untuk menyelidiki kesalahan transformasi.

Dalam makalah ini, kami menjelaskan pendekatan transformasi data, yang disebut sebagai ETL dinamis (D-ETL), yang mengotomatiskan bagian dari proses dengan menggunakan skalabel, dapat digunakan kembali

dan kode yang dapat disesuaikan, sambil mempertahankan aspek manual dari proses yang memerlukan sintaks pengkodean yang kompleks. Kontribusi dari pekerjaan ini meliputi 1) memberikan solusi praktis dan terukur untuk harmonisasi data dalam jaringan penelitian data klinis dengan sumber data yang heterogen dan 2) menurunkan hambatan teknis bagi pakar domain data kesehatan untuk memainkan peran utama dalam operasi ETL dengan menyederhanakan data proses transformasi.

Metode

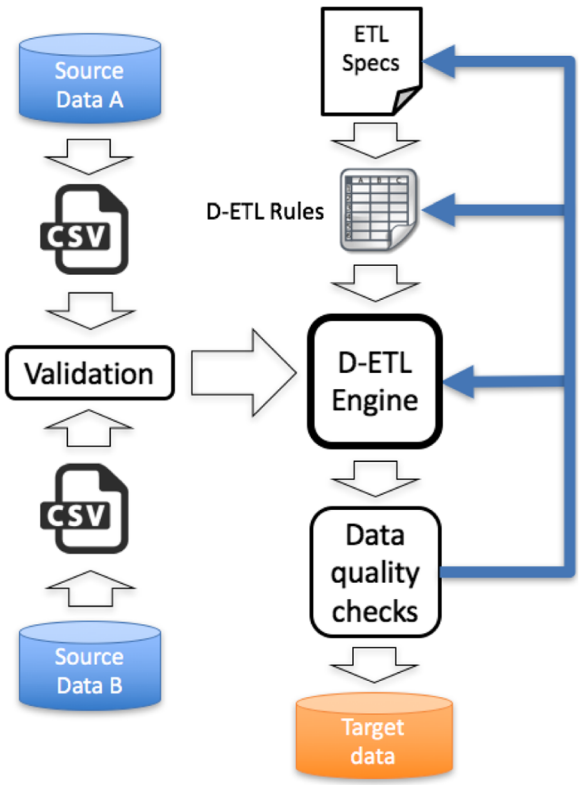
Pengaturan dan konteks

SAFTINet (Scalable Architecture for Federated Translational Inquiries Network) adalah salah satu dari tiga jaringan penelitian terdistribusi nasional yang didanai oleh Agency for Healthcare Research and Quality (AHRQ) untuk mendukung penelitian efektivitas komparatif skala luas [21]. Pada tahun 2010 SAFTINet memilih OMOP versi 4 Model Data Umum (OMOP v4 CDM) dan Terminologi sebagai pendekatannya untuk menyelaraskan dan berbagi data di semua mitra data [32, 39]. Setiap lembaga berbagi data yang berpartisipasi dalam SAFTINet harus membuat dan memelihara database yang berisi data EHR mereka yang direstrukturisasi menjadi sesuai dengan HIPAA (HIPAA = Undang-Undang Portabilitas dan Akuntabilitas Asuransi Kesehatan), kumpulan data terbatas yang sesuai dengan OMOP CDM. Data klinis juga terintegrasi dengan data klaim yang disediakan oleh pembayar, untuk dua organisasi jaring pengaman, dan data hasil yang dilaporkan pasien dikumpulkan di titik perawatan untuk membuat database umum SAFTINet. Untuk memastikan kepatuhan terhadap Aturan Privasi dan Keamanan HIPAA, SAFTINet membatasi elemen data informasi kesehatan yang dilindungi (PHI) ke elemen yang diizinkan menurut definisi peraturan dari kumpulan data terbatas (LDS), yang menghapus pengidentifikasi langsung, seperti nama, alamat, dan jaminan sosial jumlah; tetapi termasuk tanggal, kota/kota, negara bagian dan kode pos 3 digit [40-44]. Oleh karena itu, aturan D-ETL harus menerapkan pembatasan HIPAA ini sebagai bagian dari proses transformasi data.

Pendekatan D-ETL

Gambar 1 mengilustrasikan alur kerja dalam pendekatan D-ETL untuk mengintegrasikan dua kumpulan data sumber. Pendekatan D-ETL didasarkan pada empat komponen utama:

1. Spesifikasi ETL yang komprehensif, yang merupakan rencana induk untuk seluruh proses ETL, menguraikan dalam teks naratif dan diagram ruang lingkup data yang akan diekstraksi, model data target, dan format file data input dan output.
2. Aturan D-ETL disusun dalam format teks biasa, yang memastikan bahwa aturan dapat dibaca manusia dan karenanya mudah diteliti, dipelihara, dibagikan, dan digunakan kembali.
3. Mesin aturan ETL efisien yang menghasilkan pernyataan SQL lengkap dari aturan ETL untuk mengubah, menyesuaikan, dan memuat data ke tabel target.



Gambar 1 Alur kerja pendekatan D-ETL untuk mengintegrasikan dua kumpulan data sumber

4. Pernyataan SQL yang dibuat secara otomatis ini dapat diakses oleh perancang ETL untuk mengeksekusi, menguji, dan men-debug aturan sehingga mendukung proses berulang validasi dan debugging.

Spesifikasi dan desain ETL

Pedoman spesifikasi ETL (dibuat dalam aplikasi pengolah kata standar) berisi informasi tentang skema sumber dan target, pemetaan terminologi antara elemen dan nilai data dalam skema sumber dan target, serta definisi dan konvensi untuk data dalam skema target. Dokumen spesifikasi ETL dibuat oleh satu atau lebih pakar domain data kesehatan yang memiliki pengetahuan luas tentang skema sumber dan target.

Ekstraksi dan validasi data

Pada langkah ekstraksi data, elemen data yang diperlukan dari sistem sumber diekstraksi ke penyimpanan data sementara dari mana elemen tersebut diubah dan dimuat ke dalam database target. Pendekatan D-ETL menggunakan file teks commaseparated values (CSV) untuk pertukaran data karena penggunaannya yang luas dan dapat diterima [45]. Data yang diekstraksi kemudian melalui proses validasi data termasuk pemeriksaan data input untuk data yang hilang di bidang yang diperlukan dan nilai kunci asing yatim piatu (mis.

kolom kunci tetapi tidak di kolom kunci utama) centang. Selain itu, proses transformasi data biasanya memiliki asumsi khusus tentang nilai dan struktur data masukan yang memerlukan validasi. Gambar 2 menunjukkan daftar contoh aturan validasi.

Aturan D-ETL

Dengan D-ETL, data input yang divalidasi ditransformasikan melalui seperangkat aturan D-ETL. Struktur aturan D-ETL memerlukan informasi dasar tentang database sumber dan target (yaitu database, skema, tabel, bidang) serta formula transformasi data. Struktur aturan D-ETL memungkinkan data target dihasilkan dengan menggabungkan data daribeberapa tabel sumber terkait. Contoh spesifik dari masalah transformasi data yang dapat diatasi oleh aturan ETL adalah transfer data sumber dari tabel Demografi sumber dan bidang Ras ke tabel Person target dan bidang Ras di OMOP CDM. Asumsikan bahwa nilai Ras dalam data EHR sumber dikodekan menggunakan standar Health Level 7 (HL7)[3] sistem pengkodean. Karena sistem pengkodean standar untuk nilai Ras di OMOP adalah Systematized Nomenclature of Medicine (SNOMED),[4] harus ada operasi pemetaan terminologi sebagai bagian dari proses ETL. Untuk mengatasi masalah ini, aturan D-ETL yang mengubah data dalam tabel Demografi sumber harus merujuk setidaknya dua tabel sumber: Tabel Demografi dan tabel Source_to_Concept_Map. Tabel Source_to_Concept_Map menyediakan pemetaan dari kode nilai HL7 untuk balapan ke kode nilai SNOMED untuk balapan.

Aturan D-ETL adalah struktur data yang memiliki 12 atribut dan baris sebanyak yang diperlukan untuk spesifikasi aturan yang lengkap. Setiap aturan menghasilkan kode SQL yang bertanggung jawab untuk mengubah dan memuat satu atau lebih bidang dalam atunggal meja sasaran. Tabel 1 berisi daftar atribut aturan dan deskripsinya. Aturan D-ETL biasanya disusun oleh pakar domain data kesehatan berdasarkan dokumen spesifikasi ETL. Aturan D-ETL menerapkan pemetaan skema yang ditentukan dalam dokumen spesifikasi ETL.

Mengingat strukturnya, aturan D-ETL dapat disimpan dalam file berformat CSV dengan satu kolom untuk setiap atribut. Meskipun aturan D-ETL dalam format CSV dapat diedit menggunakan sebagian besar editor teks yang tersedia dengan semua sistem operasi utama, pengalaman menunjukkan bahwa aturan D-ETL dapat disusun dan dipelihara dengan baik dalam aplikasi spreadsheet. Aturan D-ETL dapat dibagikan dengan mudah di antara tim ETL yang memiliki data sumber dan struktur data target yang sama. Jika beberapa sumber data dimuat ke dalam satu kumpulan data target, setiap sumber data memiliki kumpulan aturan D-ETL sendiri. Tabel 2 adalah contoh aturan D-ETL. Untuk penyederhanaan, beberapa atribut dihilangkan dalam contoh. Menggunakan SQL, transformasi yang ditentukan dalam contoh aturan D-ETL ini dapat dilakukan dengan menggunakan sepasang pernyataan SELECT dan INSERT,[5] mengikuti sintaks di bawah ini:

```
INSERT INTO tableName <fieldList> SELECT
<Transformed fieldList> FROM <tableList>
```

Pernyataan INSERT dan SELECT di atas dihasilkan secara otomatis oleh mesin D-ETL dari aturan D-ETL. Setiap komponen aturan sesuai dengan operasi kueri tertentu. Mesin aturan D-ETL secara langsung mendukung operasi SQL berikut: INSERT, SELECT, SELECT DISTINCT, JOINS (inner join, left outer join, right outer join, full outer join) dan WHERE. Struktur aturan D-ETL memanfaatkan kesederhanaan format CSV dan fleksibilitas pernyataan SQL lengkap. Perancang D-ETL dapat membuat aturan tanpa memiliki pengetahuan yang luas tentang sintaks formal SQL, dan

Validation error type	Action to take
A column is present in a data file but not mentioned in the schema	◉ Ignore and don't load the column ○ Fail validation
Data in a date or timestamp column cannot be parsed as a date	○ Ignore and load as-is ◉ Fail validation
Data is missing in a field defined as required by the schema	○ Ignore and load the blank data ◉ Fail validation
A column in the schema has a missing length, precision or scale	◉ Ignore and use defaults for validation ○ Fail validation

Gambar 2. Contoh aturan validasi data input dengan kriteria validasi longgar dan ketat

Tabel 1 Jenis Atribut dari aturan D-ETL

Atribut	Deskripsi	Kelompok
Urutan aturan	Nomor identifikasi aturan. Semua baris aturan harus memiliki urutan aturan yang sama	Identifikasi
Deskripsi aturan	Deskripsi singkat dengan panjang maksimum 255 karakter untuk menjelaskan tujuan aturan.	Identifikasi
Basis data sasaran	Nama basis data target	Target
Skema target	Nama skema target	Target
Tabel target	Nama tabel target Nama	Target
Kolom sasaran	kolom target	Target
Jenis peta	Jenis baris. Kemungkinan nilai: PRIMARY, JOIN, LEFT JOIN, RIGHT JOIN, FULL JOIN, WHERE, VALUE, CUSTOM.	Sumber
Urutan peta	Identifikasi baris dalam aturan	Identifikasi
Basis data sumber	Nama database sumber	Sumber
Skema sumber	Nama skema sumber	Sumber
Tabel sumber	Nama tabel sumber	Sumber
Nilai sumber	VALUE baris: Nilai yang digunakan untuk mengisi kolom target GABUNG baris: gabungkan kondisi WHERE baris: kondisi di mana	Sumber

hanya memerlukan masukan dari ahli teknis untuk keadaan khusus (misalnya, debugging aturan yang kompleks). Semua aturan D-ETL yang digunakan untuk memuat satu set data terkandung dalam satu berkas CSV.

Atribut aturan D-ETL dapat menjadi dikategorikan menjadi tiga komponen fungsional: aturan spesifikasi, keluaran tujuan dan sumber data. Atribut spesifikasi aturan meliputi: urutan aturan, deskripsi aturan, dan ID sumber data, bidang yang digunakan untuk mengidentifikasi kumpulan data tertentu jika beberapa kumpulan data dengan kumpulan aturan berbeda diproses secara bersamaan. Kunci komposit yang secara unik mengidentifikasi aturan dibentuk berdasarkan kombinasi ketiga bidang ini. Urutan aturan adalah pengidentifikasi unik aturan. Namun, karena setiap aturan terdiri dari beberapa baris yang mewakili aturan itu's komponen, semua baris ini memiliki urutan aturan yang sama. Oleh karena itu, bersama dengan urutan aturan, setiap baris dalam aturan selanjutnya diidentifikasi oleh kolom urutan peta. Saya t'Penting untuk dicatat bahwa urutan aturan unik dalam satu kumpulan aturan, namun, mungkin tidak unik di seluruh kumpulan aturan yang berbeda. Atribut tujuan keluaran berisi informasi tentang tujuan target (misalnya database target, skema target, tabel target, kolom target). Aturan hanya dapat mengisi satu tabel target. Tidak semua atribut dari tabel target harus disertakan dalam aturan. Namun, nilai NULL biasanya digunakan sebagai nilai sumber untuk kolom yang tidak terisi.

Atribut data sumber termasuk informasi data sumber (misalnya database sumber, skema sumber, tabel sumber, nilai sumber). Kolom data sumber tidak hanya berisi informasi tentang lokasi sumber data tetapi juga rumus transformasi data.

Aturan contoh pada Tabel 2 digunakan untuk mengisi tabel target: Care_site. Aturan akan digunakan untuk menghasilkan satu pernyataan SQL. Baris dengan tipe peta PRIMARY mengidentifikasi tabel sumber utama dari mana data akan ditanyakan, dalam contoh ini tabel Medical_claims. Tabel utama adalah tabel yang memiliki setidaknya satu bidang yang digunakan untuk mengisi kunci utama dari tabel target. Kolom map_type dari baris pertama selalu disetel ke“UTAMA”, indikasi tabel utama dari mana sumber data berada. Tabel sumber tambahan dapat digabungkan dengan tabel utama oleh operator gabungan dengan map_type = {JOIN, LEFT JOIN, RIGHT JOIN, FULL JOIN} dan kondisi gabungan yang ditentukan dalam kolom map_type. Dalam contoh, kolom source_value dari baris PRIMARY terdiri dari primary key gabungan yang digunakan untuk mengisi primary key dari tabel target. Kunci utama mencakup 3 bidang: billing_provider_id dan place_of_ service_code dari tabel klaim medis dan provider_organization_type dari tabel provider. Tabel penyedia digabungkan dengan tabel medical_claims dalam baris GABUNG dengan kondisi GABUNG yang ditentukan dalam nilai_sumber dari baris yang sama. Baris opsional dengan klausa WHERE dapat digunakan dengan kondisi WHERE di kolom source_value. Sebagai contoh,'1'

atau '2' akan diisi ke tabel target. Perhatikan bahwa“di” operator digunakan karena merupakan operator standar yang didukung PostgreSQL. Baris berikutnya, yang memiliki VALUE sebagai map_type, berisi pemetaan langsung dari nilai sumber ke kolom target. Nilai NULL harus ditunjukkan dengan jelas di mana kolom target tidak dapat diisi. Meskipun semua baris dalam satu aturan memiliki aturan_urutan dan deskripsi aturan yang sama, mereka akan memiliki urutan_peta yang berbeda. Terakhir, setiap nilai yang tercantum dalam kolom source_value harus menyertakan tabel sumber dan bidang sumber.

Aturan D-ETL dapat mencakup pemetaan yang memiliki tingkat kompleksitas berbeda yang bervariasi dari pemetaan satu-ke-satu

Meja 2 Contoh aturan D-ETL yang memuat data ke tabel Care_site di OMOP dari file CSV sumber berbasis klaim

Urutan Aturan	Deskripsi Aturan	Tabel Target	Kolom Target	Jenis peta	Urutan Peta	Tabel Sumber	Nilai Sumber
1	Medical_claims ke Care_site	Care_site		UTAMA	1	Medical_claims	medical_claims.billing_provider_id, medical_claims.place_of_service_code, provider.provider_organization_type
1	Medical_claims ke Care_site	Care_site		IKUTI	2	pemberi	medical_claims.billing_provider_id = penyedia.provider_id
1	Medical_claims ke Care_site	Care_site		DIMANA	3		provider.provider_organization_type di ('1', '2')
1	Medical_claims to Care_site	Care_site	care_site_source_value	NILAI	4		medical_claims.billing_provider_id || '-' ||medical_claims.place_of_service_code||'-' ||provider.provider_organization_type
1	Medical_klaim ke Care_site	Care_site	organization_source_value	NILAI	5		BATAL
1	Medical_claims ke Care_site	Care_site	place_of_service_source_value	NILAI	6		medical_claims.place_of_service_code
1	Medical_claims ke Care_site	Care_site	care_site_address_1	NILAI	7		provider.provider_address_first_line
1	Medical_claims ke Care_site	Care_site	care_site_address_2	NILAI	8		provider.provider_street
1	Medical_claims ke Care_site	Care_site	care_site_city	NILAI	9		provider.provider_city
1	Medical_claims ke Care_site	Care_site	care_site_state	NILAI	10		provider.provider_state
1	Medical_claims ke Care_site	Care_site	care_site_zip	NILAI	11		provider.provider_zip
1	Medical_claims ke Care_site	Care_site	care_site_county	NILAI	12		BATAL

untuk pemetaan banyak-ke-satu. Output dari aturan dapat difilter oleh klausa WHERE yang ditentukan dalam kolom map_type. Baris WHERE dalam aturan D-ETL bukanlah komponen yang diperlukan, tetapi jika ada, setiap aturan hanya dapat memiliki satu klausa WHERE. Kolom source_value juga dapat berisi ekspresi yang dapat diformulasikan menggunakan dialek asli DBMS untuk mengubah data dari sumber ke target. Misalnya, jika PostgreSQL adalah DBMS, semua operator dan fungsi PostgreSQL didukung. Pendekatan ini memperluas fleksibilitas pendekatan hybrid dengan memungkinkan desainer D-ETL untuk mengambil keuntungan dari semua fungsi yang didukung oleh DBMS target. Kelemahan dari pendekatan ini adalah bahwa terjemahan kode diperlukan ketika aturan digunakan dalam DBMS yang berbeda. Oleh karena itu, adalah praktik yang baik untuk menggunakan fungsi SQL yang banyak digunakan jika memungkinkan.

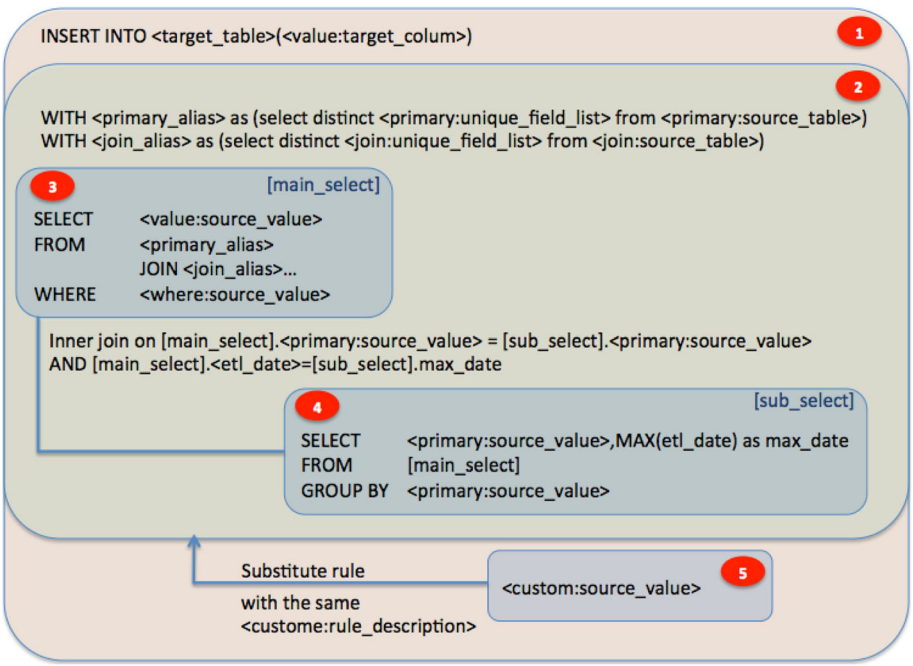
Mesin D-ETL

Aturan D-ETL disusun dalam format yang dapat dibaca manusia, memungkinkan personel dengan keahlian pemrograman basis data terbatas untuk menyusun, membaca, memverifikasi, dan memeliharanya. Mesin D-ETL secara otomatis menerjemahkan aturan D-ETL ini ke dalam pernyataan SQL kompleks yang dapat dieksekusi untuk mengubah dan memuat data ke dalam tabel target. Selama proses ini, beberapa peningkatan kinerja kueri dan operasi pembersihan data seperti ekspresi tabel umum (CTE) dan de-duplikasi data secara otomatis dimasukkan ke dalam pernyataan SQL. Mesin D-ETL terdiri dari lima sub-proses yang berhubungan dengan integrasi data, transformasi, de-duplikasi dan memuat. Pada Gambar 3, angka dalam oval merah mewakili proses yang digunakan untuk menjalankan setiap subproses mesin D-ETL. Dalam diagram, variabel diapit oleh tanda kurung siku (<variable>) dan kueri diapit dalam tanda kurung siku ([query]). Atribut aturan D-ETL diidentifikasi dengan format: <map_type:column_name>. Misalnya, kondisi WHERE aturan dapat diidentifikasi dengan <where:source_value>. Meskipun urutan eksekusi berbeda, untuk keterbacaan, proses akan dihitung dari atas ke bawah.

Dalam proses 1, mesin D-ETL memulai dengan membuat pernyataan INSERT menggunakan nilai di kolom target. Dalam proses 2, data diduplikasi menggunakan subquery dan CTE. Dalam proses 3, kumpulan data sumber yang tidak terduplikasi kemudian digabungkan dan disaring sebelum dimasukkan ke dalam tabel target dalam proses 1. Contoh solusi untuk catatan yang bentrok adalah dengan memilih catatan dengan stempel waktu ETL terbaru dari catatan grup yang memiliki kesamaan. nilai kunci utama. Kueri kustom ditangani dalam proses 5. Lihat File tambahan 1 untuk deskripsi aturan CUSTOM dan File tambahan 2 untuk deskripsi detail proses mesin D-ETL individual.

Pengujian dan debugging

Pendekatan D-ETL memfasilitasi pengujian dan debugging tingkat kode. Pernyataan SQL yang dihasilkan oleh mesin D-ETL disimpan secara terpisah dalam database dari aturan individu. Desainer D-ETL dapat menguji dan men-debug pernyataan ini dan meninjau hasil kueri secara langsung dalam proses pengujian internal. Setiap kesalahan sintaksis dan semantik di



Gambar 3 Arsitektur mesin ETL

pernyataan dapat ditelusuri kembali ke komponen aturan. Mekanisme ini memungkinkan desainer D-ETL untuk memahami kesalahan pada tingkat pernyataan SQL dan memanfaatkan sistem pesan kesalahan DBMS, daripada harus melalui pesan kesalahan tidak langsung yang disediakan oleh GUI atau file log kesalahan setelah ETL lengkap. proses.

Selain itu, mengeksekusi pernyataan SQL secara langsung memungkinkan desainer D-ETL untuk secara iteratif melakukan penyesuaian coba-coba pada kueri hingga transformasi yang diinginkan dibuat alih-alih terus mengubah aturan D-ETL. Akibatnya, hanya perubahan akhir yang perlu dicatat dalam spesifikasi D-ETL. Mampu meninjau pernyataan SQL secara langsung memungkinkan desainer D-ETL untuk memahami hubungan antara desain aturan dan hasil kueri, oleh karena itu, meningkatkan proses desain aturan. Kelemahan dari proses pengujian dan debugging ini adalah membutuhkan akses langsung ke database backend tempat pernyataan D-ETL disimpan dan mungkin memerlukan pengetahuan SQL tingkat lanjut. Tabel 3 merangkum tantangan proses ETL dan solusi untuk tantangan ini yang dimungkinkan oleh D-ETL.

Hasil

Pengujian dan validasi internal

Mitra SAFTINet berhasil mengimplementasikan D-ETL melalui sistem yang disebut ROSITA (Reusable OMOP and SAFTINet Interface Adaptor). Sistem perangkat lunak ROSITA dioperasikan melalui antarmuka pengguna berbasis web yang didukung oleh mesin D-ETL backend. Efisiensi D-ETL menggunakan ROSITA untuk memproses data kesehatan sangat menjanjikan, bahkan dalam situasi di mana ada data duplikat dan tumpang tindih dalam kumpulan data sumber. Tabel 4 menunjukkan runtime dari beberapa D-ETL

Tabel 3 Tantangan dan solusi

Tantangan	Solusi dengan Pendekatan D-ETL
Heterogenitas dalam kumpulan data sumber	• Spesifikasi ETL • Mesin D-ETL berbasis aturan • Penerimaan kode SQL asli • Mekanisme aturan khusus
Ekstraksi data mengganggu sumber EHR	• format file CSV
Efisiensi	• Mesin D-ETL terintegrasi • Pengoptimalan kueri
Duplikat dan tumpang tindih data	• De-duplikasi data otomatis dan pemuatan data tambahan
Kualitas data	• Data input: Validasi data yang diekstraksi • Data keluaran: Pembuatan profil dan visualisasi data
Keahlian manusia	• Struktur aturan eksplisit • Pengujian aturan dan mekanisme debugging yang efektif
Resumption (kemampuan untuk melanjutkan dari titik di mana kesalahan sebelumnya terjadi)	• Proses ETL modular

Tabel 4 Performa mesin D-ETL di ROSITA

Aturan jumlah	Jumlah tabel sumber	Jumlah catatan (di semua tabel sumber)	Waktu tayang (dalam detik)
1	1	21.565	1.1
2	2	851.706	30.3
3	2	1.910.513	12.0
4	2	1.324.860	13.1
5	3	1.987.582	15.3
6	3	2,007.661	30.1

aturan dalam ROSITA dari proses pengujian internal, memuat set data kesehatan pada mesin virtual CentOS 6 dengan 2 inti pemrosesan, RAM 32GB, dan hard drive 300GB.

Implementasi D-ETL yang praktis, terukur, dan layak di ROSITA

Menggunakan ROSITA, mitra SAFTINet di ketiga negara bagian dapat memuat data klinis dan mengubah data menjadi OMOP v4 CDM dan di dua negara bagian ini, di mana data klaim tersedia, mitra dapat memuat dan menautkan data klinis dan klaim, sebelum untuk menyelesaikan transformasi data ke OMOP. Dua mitra dengan staf yang baik, departemen informatika yang canggih menerapkan ROSITA di lingkungan mereka sendiri, memetakan langsung dari EHR atau gudang data elektronik (EDW) mereka sendiri ke OMOP. Sepuluh mitra lainnya menggunakan perantara yang merupakan pakar domain data kesehatan tetapi bukan pemrogram database SQL tingkat lanjut untuk mengubah data EHR mereka menjadi model data perantara; mereka kemudian menerapkan D-ETL dalam ROSITA untuk mengubah data dari model perantara menjadi OMOP. Empat instans ROSITA yang berafiliasi dengan SAFTINet yang dihasilkan saat ini berisi catatan untuk 1.616.868 nyawa pasien. Sistem ROSITA ini juga telah digunakan untuk menangkap hasil lebih dari 8000 ukuran hasil yang dilaporkan pasien, yang digunakan dalam studi SAFTINet CER [46, 47].

Panduan spesifikasi ETL khusus situs digunakan oleh setiap mitra data untuk memandu dan mendokumentasikan pilihan data sumber yang diekstraksi dan lokasi target yang diinginkan. Data sumber diekstraksi dan ditransfer menggunakan file CSV. Dalam praktiknya, format CSV adalah penyimpanan sementara yang fleksibel dan ada di mana-mana untuk D-ETL. Mengekstrak data ke file CSV memungkinkan proses ekstraksi data dipisahkan dari proses transformasi dan pemuatan data yang lebih sulit. Memisahkan ekstraksi data dari transformasi data menghilangkan kebutuhan akan koneksi jaringan aktif ke data sumber setiap kali tugas transformasi baru dilakukan. Selain itu, mengekstrak data sumber ke dalam sistem penyimpanan sementara, tanpa terhubung langsung ke database sumber, memungkinkan akses terkontrol ke data yang diekspor yang dibuat oleh pemilik data.

Validasi kualitas data pada kumpulan data yang diekstraksi penting untuk memastikan keberhasilan langkah selanjutnya. Validasi data biasanya terjadi segera setelah langkah ekstraksi data untuk identifikasi cepat masalah apa pun dengan data yang diekstraksi dan ekstraksi ulang yang dipercepat. Pengalaman menunjukkan bahwa sangat sulit untuk menghasilkan kumpulan data yang diekstraksi sempurna yang akan diterima oleh proses ETL pada percobaan pertama. Untuk alasan itu, penting untuk belajar dari kesalahan dan memasukkannya kembali ke dalam dokumen konvensi ekstraksi data untuk ekstraksi di masa mendatang. Data dalam file CSV dapat divalidasi secara langsung atau diimpor apa adanya ke dalam DBMS, idealnya DBMS yang sama tempat transformasi akan terjadi. Tabel 5 mencakup daftar aturan validasi data yang dilakukan untuk memastikan kualitas data sesuai untuk proses ETL.

Pada Tabel 5, jika jenis kesalahannya adalah “Kesalahan”, data gagal aturan validasi dan harus diekstraksi lagi setelah masalah diperbaiki. Elemen data yang menyebabkan kesalahan dan lokasi persisnya dalam kumpulan data harus disediakan. Untuk melindungi data sensitif agar tidak terekspos, hanya nomor baris rekaman data dalam file data yang ditampilkan. Jika jenis kesalahannya adalah“Peringatan”, data tidak akan gagal dalam aturan validasi. Sebagai gantinya, akan ada pesan peringatan yang memberikan informasi tentang data tersebut. Keputusan untuk menangani masalah ini adalah opsional. Daftar aturan validasi didasarkan pada antisipasi kebutuhan proses transformasi data menjadi model data target. Jenis kesalahan dapat diklasifikasikan berdasarkan dampak kesalahan dan jumlah kehilangan informasi yang diharapkan.

Di SAFTINet, pakar domain data kesehatan, yang memiliki keahlian terbatas tentang SQL atau DBMS, membuat kumpulan aturan D-ETL terpisah untuk setiap mitra's OMOP CDM instantiasi. Kadang-kadang, pakar domain memerlukan bantuan dari personel teknis untuk fungsi DBMS yang kompleks. Bantuan teknis juga diperlukan jika terjadi perbedaan keluaran data yang tidak jelas. Seiring waktu, jumlah bantuan teknis mungkin berkurang karena pengalaman pakar domain Dalam banyak kasus, pakar domain data kesehatan dapat menyusun, memuat, dan men-debug aturan. Melalui operasi ROSITA, kami menemukan bahwa D-ETL sangat efektif dalam pengujian aturan dan debug. Pakar domain data kesehatan mampu melacak sumber kesalahan secara efektif dengan menjalankan aturan individual.

Meluas di luar SAFTINet, ROSITA telah digunakan oleh DARTNet Institute (DARTNet) untuk berhasil mengubah data lebih dari 7 juta pasien menjadi OMOP CDM untuk berbagai CER, peningkatan kualitas, dan kegiatan penelitian intervensi. DARTNet menggunakan ROSITA dengan cara yang berbeda dari mitra SAFTINet; kontributor data mengirim data yang teridentifikasi sepenuhnya ke personel DARTNet (di bawah HIPAA BAA) yang kemudian melakukan transformasi secara terpusat (yaitu, dengan cara yang tidak terdistribusi).

Diskusi

Dalam proyek ini, kami merancang dan menerapkan pendekatan ETL berbasis aturan hybrid baru yang disebut Dynamic-ETL (D-ETL). Implementasi dalam praktik menunjukkan bahwa D-ETL dan implementasinya di ROSITA adalah pendekatan yang layak dan berhasil untuk harmonisasi struktural dan sematik data kesehatan di jaringan berbagi data kesehatan besar yang berisi mitra data yang heterogen, seperti SAFTINet dan DARTNet. D-ETL memungkinkan pakar domain data kesehatan dengan keahlian SQL terbatas untuk terlibat dalam semua fase, yaitu spesifikasi ETL, desain aturan, aturan pengujian dan debugging, dan hanya memerlukan bantuan teknis ahli dalam kasus khusus. D-ETL mempromosikan struktur aturan yang mengakomodasi operasi ETL langsung dan kompleks dan mendukung transparansi ETL dan mendorong pembagian aturan D-ETL. Mesin aturan ETL juga menggabungkan mekanisme yang menangani data yang saling bertentangan dan duplikat. Menggunakan perangkat keras yang tersedia, sistem D-ETL yang diimplementasikan menunjukkan hasil kinerja yang dapat diterima yang memuat data kesehatan nyata.

Keakuratan dan keandalan aturan D-ETL dan mesin D-ETL bergantung pada keakuratan dan keandalan konten aturan D-ETL. Metrik kinerja teknis tambahan untuk diperiksa dalam implementasi skala yang lebih besar akan mencakup aspek kualitas data yang ditingkatkan, termasuk akurasi (misalnya, lebih sedikit kesalahan dalam pemetaan) dan kelengkapan (misalnya, lebih sedikit titik data yang hilang) [48]. Faktor kunci lainnya dalam adopsi dan penggunaan D-ETL termasuk kegunaan dan kegunaan yang dirasakan dan sumber daya dan waktu yang dibutuhkan untuk mengimplementasikan sistem,

Tabel 5 Contoh validasi data yang diekstraksi

Aturan Validasi	Tipe kesalahan	Deskripsi
Data dalam kolom tanggal atau stempel waktu tidak dapat diuraikan sebagai tanggal	Kesalahan	Data tanggal yang tidak valid akan menggagalkan operator dan fungsi tanggal
Data tidak ada dalam bidang yang ditentukan seperti yang diperlukan oleh skema	Kesalahan	Data yang hilang di bidang yang diperlukan akan melanggar batasan basis data skema target
Kolom dalam skema memiliki panjang, presisi, atau skala yang hilang	Peringatan	Panjang default, presisi atau skala dapat digunakan
Data dalam kolom numerik atau desimal bukan angka Data terlalu	Kesalahan	Data numerik yang tidak valid akan menggagalkan operator dan fungsi numerik
panjang untuk bidang teks atau varchar	Kesalahan	Kehilangan data akan terjadi jika nilai teks panjang dipotong untuk memenuhi persyaratan panjang

yang dapat dinilai dengan menggunakan survei dan wawancara mendalam dengan pengguna [49]. Dalam pengembangan awal dan pekerjaan implementasi skala kecil ini, pengguna mitra klinis melaporkan ahli domain yang mengikuti pendekatan DETL mengalami kurva pembelajaran dan peningkatan efisiensi setelah penggunaan pertama. Pemahaman yang lebih baik tentang upaya yang terkait dengan komposisi dan debugging aturan D-ETL, dan tingkat keterlibatan teknisi database dibandingkan dengan alat atau pendekatan lain akan lebih memvalidasi efektivitas dan efisiensi relatif D-ETL dalam konteks yang berbeda. Penting untuk dicatat bahwa kinerja operasi ETL dengan data perawatan kesehatan bergantung pada banyak faktor yang saling terkait seperti keakraban dengan model data sumber dan kerumitan proses pemetaan terminologi yang berada di luar pengoperasian mesin D-ETL.

Terlepas dari kelebihannya, D-ETL memiliki beberapa keterbatasan. Pertama, meskipun keahlian dalam penulisan kueri tidak diperlukan, keterampilan pengkodean SQL tertentu diperlukan untuk pakar domain data kesehatan yang terlibat dalam proses ETL. Pengetahuan tentang operator dan fungsi DBMS diperlukan untuk pembuatan aturan. Kedua, karena aturan ETL disusun dalam perangkat lunak pihak ketiga seperti Excel, pemeriksaan kesalahan sintaksis waktu nyata tidak tersedia. Penyusun aturan tidak akan tahu tentang kesalahan sintaksis (yaitu nama kolom yang salah) sampai pernyataan SQL dibuat. Ketiga, proses pengujian dan debugging memerlukan akses langsung ke database aturan dan mengekstrak dataset, yang mungkin tidak tersedia untuk penyusun aturan karena keterbatasan akses keamanan database.

Arah masa depan D-ETL fokus pada mengatasi beberapa keterbatasan dan meningkatkan efisiensi proses perancangan aturan. Pertama, munculnya metode pemetaan skema semiotomatis mendukung otomatisasi komposisi aturan D-ETL [50]. Pelibatan pakar domain data kesehatan selanjutnya dapat lebih fokus pada perbaikan hasil pemetaan dan memastikan kualitas data. Kedua, mekanisme validasi aturan otomatis yang memeriksa kesalahan sintaksis dasar akan meningkatkan efisiensi proses pembuatan aturan ETL. Agar layak, editor aturan yang ramah pengguna dengan antarmuka pengguna yang intuitif harus dikembangkan. Ketiga, ekspresi dalam aturan ETL harus dalam bahasa DBMS lokal. Untuk aturan yang akan digunakan dan digunakan kembali di DBMS yang berbeda, alat konversi aturan yang secara otomatis menerjemahkan operator dan fungsi dari satu dialek SQL ke yang lain diperlukan. Alat sumber terbuka, seperti SQL Renderer dari komunitas OHDSI,[6] bisa menjadi solusi potensial untuk masalah ini. Akhirnya, meskipun aturan disusun dalam format teks biasa, presentasi grafis dari struktur aturan akan meningkatkan pemeliharaan aturan ETL dan membantu anggota tim ETL memahami aturan kompleks yang dibuat oleh orang lain.

Kesimpulan

Harmonisasi data merupakan langkah penting menuju interoperabilitas data yang mendukung kemajuan penelitian efektivitas komparatif. Harmonisasi data dapat dicapai dengan memasukkan standar data, pengetahuan ahli domain dan proses dan alat ETL yang efektif dan efisien. Dalam pekerjaan ini, kami mengusulkan pendekatan ETL data dinamis untuk menurunkan hambatan teknis yang dihadapi selama pelaksanaan proses ETL. Pendekatan kami telah diterapkan dan digunakan untuk memuat data klinis dan klaim dari sistem catatan kesehatan elektronik sumber ke dalam model data umum OMOP. Ini adalah langkah maju yang penting untuk membuat data berkualitas tinggi tersedia untuk peningkatan kualitas klinis dan penelitian biomedis.

Catatan akhir

[1] https://www.talend.com/
[2] http://www.pentaho.com/
[3] http://www.hl7.org/
[4] https://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html
[5] http://www.w3schools.com/sql/sql_insert.asp
[6] https://github.com/OHDSI/SqlRender

File tambahan

Berkas tambahan 1: Deskripsi aturan D-ETL khusus. (DOCX 13 kb)
Berkas tambahan 2: Deskripsi mesin D-ETL. (DOCX 21 kb)

Singkatan

AHRQ: Badan Penelitian dan Mutu Kesehatan; BAA: Perjanjian asosiasi bisnis; CDM: Model data umum; CER: Penelitian efektivitas komparatif; CSV: nilai yang dipisahkan koma; DARTNet: Penelitian rawat jalan terdistribusi dalam jaringan terapi; DBMS: Sistem manajemen basis data; DETL: Ekstraksi, transformasi, dan pemuatan dinamis; EDW: Gudang data elektronik; EHRs: Catatan kesehatan elektronik; ETL: Ekstraksi, transformasi, dan pemuatan; GUI: Antarmuka pengguna grafis; HIPAA: Tindakan portabilitas dan akuntabilitas asuransi kesehatan; HL7: Kesehatan tingkat 7; I2b2: Informatika untuk mengintegrasikan biologi dan tempat tidur; ICD-9CM: Klasifikasi penyakit internasional, revisi kesembilan, modifikasi klinis; LDS: Kumpulan data terbatas; MS: Mini-Sentinel; OHDSI: Ilmu data kesehatan observasi dan informatika;
OMOP: Kemitraan hasil medis observasional; PCOR: Penelitian hasil yang berpusat pada pasien; PCORnet: Jaringan penelitian hasil yang berpusat pada pasien; PHI: Informasi kesehatan yang dilindungi; ROSITA: Adaptor Antarmuka OMOP dan SAFTINet yang dapat digunakan kembali; SAFTINet: Arsitektur yang dapat diskalakan untuk jaringan pertanyaan translasi gabungan; SNOMED: Tata nama obat yang sistematis; SQL: Bahasa kueri terstruktur

Ucapan Terima Kasih

Mr David Newton memberikan dukungan pemrograman untuk antarmuka pengguna grafis dari sistem. Kami berterima kasih kepada anggota tim di Recombinant by Deloitte yang telah mendukung pengembangan sistem ROSITA.

Pendanaan

Pendanaan disediakan oleh AHRQ 1R01HS019908 (Scalable Architecture for Federated Translational Inquiries Network) dan AHRQ R01 HS022956 (SAFTINet: Mengoptimalkan Nilai dan Mencapai Keberlanjutan)' tanggung jawab tunggal dan tidak harus mewakili AHRQ atau NIH resmi.

Ketersediaan data dan bahan

Tidak ada dataset khusus yang dibahas dalam makalah ini. Kumpulan data yang digunakan untuk mengukur kinerja pemuatan sistem ROSITA adalah data nyata dan tidak dapat dipublikasikan karena undang-undang HIPAA dan peraturan federal, negara bagian, dan institusional lainnya.
Tautan ke kode sumber ROSITA: https://github.com/SAFTINet/rosita2-1

Implementasi alat ekstraksi data

Alat ekstraksi data diimplementasikan oleh pemilik data SAFTInet's situs. Proyek SAFTINet menerima persetujuan dari Dewan Peninjau Internal untuk infrastruktur implementasinya dan memiliki perjanjian asosiasi bisnis HIPAA untuk pengujian dengan data nyata.

Pengarang' kontribusi

Semua penulis terlibat dalam konsepsi dan desain penelitian ini. LS dan MK mengawasi persyaratan dan desain pendekatan D-ETL. TO dan CU mengimplementasikan dan mengoptimalkan mesin D-ETL. PH dan TY menyusun aturan DETL dan menguji mesin D-ETL. EB melakukan pemetaan terminologi dan menguji kualitas pemetaan terminologi. BK mengoordinasikan proses pengembangan ROSITA dan memberikan masukan penting tentang tata letak naskah. TO, LS, BK dan MK menyusun naskah tersebut. Semua penulis berpartisipasi dalam revisi naskah dan menyetujui versi final.

Persetujuan etika dan persetujuan untuk berpartisipasi

Pekerjaan yang dijelaskan dalam makalah ini menjelaskan pengembangan, pengujian, dan implementasi infrastruktur ETL, tetapi tidak menjelaskan penelitian apa pun yang menggunakan infrastruktur ini. Alat ETL diimplementasikan di belakang pemilik data' firewall, atau di belakang firewall mereka yang memiliki HIPAA Business Associate Agreements (BAA) dengan pemilik data. Persetujuan etis dan hukum untuk pekerjaan ini adalah sebagai berikut: 1. IRB menyetujui dan" protokol infrastruktur", menunjukkan bahwa infrastruktur adalah mekanisme yang disetujui untuk memproduksi dan mentransmisikan kumpulan data terbatas HIPAA untuk penelitian masa depan. 2. Setiap data yang dirilis untuk penelitian akan memerlukan persetujuan IRB terpisah dan perjanjian penggunaan data HIPAA dengan pemilik data. 3. Pemilik data menandatangani dokumen hukum ("Perjanjian Konsorsium Utama") dikembangkan bekerja sama dengan penasihat hukum universitas dan setiap organisasi' penasihat hukum sendiri, yang menentukan kondisi di mana alat ETL diimplementasikan, dipelihara, dan digunakan untuk menghasilkan kumpulan data penelitian dalam jaringan penelitian. 4. Untuk pengembangan dan pengujian, beberapa pemilik data memberikan data pasien yang teridentifikasi sepenuhnya kepada pengembang perangkat lunak, di bawah BAA HIPAA antara pemilik data dan universitas (juga mencakup kontraktor). Data ini tidak digunakan untuk penelitian. Proses tata kelola data ini dibahas dengan sangat rinci dan ditetapkan seperti yang dijelaskan dengan pejabat hukum dan regulator dari universitas dan pemilik data.

Persetujuan untuk publikasi

Tak dapat diterapkan.

Kepentingan bersaing

Para penulis menyatakan bahwa mereka tidak memiliki kepentingan yang bersaing.

Penerbit's Catatan

Springer Nature tetap netral sehubungan dengan klaim yurisdiksi dalam peta yang diterbitkan dan afiliasi institusional.

Detil penulis

[1]Departments of Pediatrics, University of Colorado Anschutz Medical Campus, School of Medicine, Building AO1 Room L15-1414, 12631 East 17th Avenue, Mail Stop F563, Aurora, CO 80045, USA. [2]Departemen Kedokteran Keluarga, Kampus Medis Universitas Colorado Anschutz, Fakultas Kedokteran, Aurora, CO, AS. [3]Departemen Kedokteran, Kampus Medis Universitas Colorado Anschutz, Fakultas Kedokteran, Aurora, CO, AS. [4] Institut Ilmu Klinis dan Penerjemahan Colorado, Kampus Medis Universitas Colorado Anschutz, Fakultas Kedokteran, Aurora, CO, AS. [5]Institut DARTNet, Aurora, CO, AS. [6]OSR Data Corporation, Lincoln, MA, AS.

Diterima: 26 Mei 2016 Diterima: 31 Agustus 2017
Published online: 13 September 2017

Referensi

1. Sox HC, Greenfield S. Penelitian efektivitas komparatif: laporan dari institut kedokteran. Ann Intern Med. 2009;151:203-5. doi:10.7326/00034819-151-3-200908040-00125.
2. Danaei G, Rodríguez LAG, Cantero OF, dkk. Data observasional untuk penelitian efektivitas komparatif: sebuah emulasi uji coba statin secara acak dan pencegahan primer penyakit jantung koroner. Metode Stat Med Res. 2013;22:70-96. doi:10.1177/0962280211403603.
3. Komite Institut Kedokteran (AS) untuk Meningkatkan Catatan Pasien. Dalam: Dick RS, Steen EB, editor. Catatan pasien berbasis komputer: teknologi penting untuk perawatan kesehatan, edisi revisi. Washington, DC: Pers Akademi Nasional; 1997.
4. Grossmann C, Institut Kedokteran (AS). Meja Bundar tentang Perawatan Kesehatan Berbasis Nilai & Sains. Data klinis sebagai bahan pokok pembelajaran kesehatan : menciptakan dan melindungi barang publik : ringkasan lokakarya. Washington, D C.: Pers Akademi Nasional; 2010. http://www.ncbi.nlm.nih.gov/bookshelf/ br.fcgi?book=nap12212 http://www.nap.edu/catalog.php?record_id=12212 Selby JV, Lipstein
5. SH. PCORI pada usia 3 tahun-kemajuan, pelajaran, dan rencana. N Engl J Med. 2014;370:592-5. doi:10.1056/NEJMp1313061.
6. Fleurence RL, Curtis LH, California RM, dkk. Meluncurkan PCORnet, jaringan penelitian klinis nasional yang berpusat pada pasien. J Am Med Menginformasikan Assoc JAMIA. 2014;21:578-82. doi:10.1136/amiajnl-2014-002747.
7. Holve E, Segal C, Lopez MH, dkk. Forum Metode Data Elektronik (EDM) untuk Penelitian Efektivitas Komparatif (CER). Perawatan Medis. 2012; 50 (Suppl): S7-10. doi:10.1097/MLR.0b013e318257a66b.
8. Rassen JA, Schneeweiss S. Menggunakan skor kecenderungan dimensi tinggi untuk mengotomatisasi kontrol pengganggu dalam sistem pengawasan keamanan produk medis terdistribusi. Farmakoepidemiol Obat Saf. 2012;21(Suppl 1):41-9. doi:10.1002/pds.2328.
9. Garbe E, Kloss S, Suling M, dkk. Skor kecenderungan dimensi tinggi versus konvensional dalam studi efektivitas komparatif coxib dan pengurangan komplikasi gastrointestinal bagian atas. Eur J Clin Pharmacol Diterbitkan Online Pertama: 5 Juli 2012. doi:10.1007/s00228-012-1334-2. Fleischer NL, Fernald LC,
10. Hubbard AE. Memperkirakan dampak potensial dari intervensi dari data observasional: metode untuk memperkirakan risiko yang dapat diatribusikan kausal dalam analisis cross-sectional gejala depresi di Amerika Latin. J Epidemiol Kesehatan Masyarakat. 2010;64:16-21. doi:10.1136/jech.2008.085985.
11. Polsky D, Eremina D, Hess G, dkk. Pentingnya variabel klinis dalam analisis komparatif menggunakan pencocokan skor kecenderungan: kasus biaya ESA untuk pengobatan anemia akibat kemoterapi. Ekonomi Farmako. 2009;27:755-65. doi:10.2165/11313860-000000000-00000.
12. McCandless LC, Gustafson P, Austin PC. Analisis skor kecenderungan Bayesian untuk data observasional. Stat Med. 2009;28:94-112. 10.1002/sim.3460. Jansen RG, Wiertz
13. LF, Meyer ES, dkk. Analisis reliabilitas data observasi: masalah, solusi, dan implementasi perangkat lunak. Metode Res Perilaku Instrum Comput. 2003;35:391-9.
14. Komite Metodologi PCORI. Standar Metodologi PCORI. 2012. http://www.pcori.org/assets/PCORI-Methodology-Standards.pdf
15. Collins FS, Hudson KL, Briggs JP, dkk. PCORnet: mengubah mimpi menjadi kenyataan. J Am Med Menginformasikan Assoc JAMIA. 2014;21:576-7. doi:10.1136/amiajnl-2014-002864.
16. Forrest CB, Margolis PA, Bailey LC, dkk. PEDSnet: Sistem Kesehatan Pembelajaran Anak Nasional. J Am Med Inform Assoc JAMIA Diterbitkan Online Pertama: 12 Mei 2014. doi:10.1136/amiajnl-2014-002743.
17. Behrman RE, Benner JS, Brown JS, dkk. Mengembangkan sistem penjaga—Sumber daya nasional untuk pengembangan bukti. N Engl J Med. 2011;364:498-9. doi:10.1056/NEJMp1014427.
18. Joe V Selby, Richard Platt, Rachael Fleurence, dkk. PCORnet: pencapaian dan rencana menuju transformasi penelitian kesehatan | PCORI. http://www.pcori.org/blog/pcornet-achievements-and-plans-road-transforming-healthresearch. Diakses 16 Mei 2016.
19. Gini R, Schuemie M, Brown J, dkk. Ekstraksi dan manajemen data dalam jaringan database perawatan kesehatan observasional untuk penelitian ilmiah: perbandingan antara strategi EU-ADR, OMOP, mini-sentinel dan MATRICE. EGEMS. 2016;4:1189. doi:10.13063/2327-9214.1189.
20. Kahn MG, Batson D, Schilling LM. Pertimbangan model data untuk peneliti efektivitas klinis. Perawatan Medis. 2012;50(Suppl):S60-7. doi:10.1097/MLR.0b013e318259bff4.

21. Schilling LM, Kwan B, Drolshagen C, dkk. Arsitektur Scalable untuk Jaringan Penerjemahan Federated Inquires (SAFTINet): infrastruktur teknologi untuk Jaringan Data Terdistribusi. EGEM Gen. jelas. Metode Meningkatkan Hasil Pasien. 2013. dalam pers.
22. Ohno-Machado L, Agha Z, Bell DS, dkk. pSCANNER: Jaringan Nasional Scalable untuk Penelitian Efektivitas yang berpusat pada pasien. J Am Med Menginformasikan Assoc JAMIA. 2014;21:621-6. doi:10.1136/amiajnl-2014-002751.
23. Murphy SN, Weber G, Mendis M, dkk. Melayani perusahaan dan seterusnya dengan informatika untuk mengintegrasikan biologi dan samping tempat tidur (i2b2). J Am Med Menginformasikan Assoc JAMIA. 2010;17:124-30. doi:10.1136/jamia.2009.000893.
24. Brown JS, Lane K, Moore K, dkk. Mendefinisikan dan mengevaluasi model database yang mungkin untuk mengimplementasikan inisiatif FDA Sentinel. 2009.https://www.brookings.edu/wp-content/uploads/2012/04/03_Brown.pdf. Diakses 15 Februari 2014.
25. Patil PS. Mekanisme ekstraksi yang ditingkatkan dalam proses ETL untuk membangun gudang data.INFLIBNET Diterbitkan Online Pertama: 22 Februari 2013. http://ir.inflibnet.ac.in:8080/jspui/handle/10603/7023. Diakses 13 November 2014.
26. Devine E, Capurro D, van Eaton E, dkk. Mempersiapkan data klinis elektronik untuk penelitian peningkatan kualitas dan efektivitas komparatif: proyek otomasi dan validasi SCOAP CERTAIN. EGEMS. 2013;1:1025. doi:10.13063/2327-9214.1025.
27. Suresh S, Gautam JP, Pancha G, dkk. Metode dan arsitektur untuk optimasi otomatis dari throughput ETL dalam aplikasi data warehousing. 2001. http://google.com/patents/US6208990. Diakses 13 November 2014.
28. Kushanoor A, Murali Krishna S, Vidya Sagar Reddy T. ETL pemodelan proses di DWH menggunakan teknik peningkatan kualitas. Teori Basis Data Int J Appl. 2013;6:179-98.
29. Peek N, Holmes JH, Sun J. Tantangan teknis untuk data besar dalam biomedis dan kesehatan: sumber data, infrastruktur, dan analitik. Tahun Med Menginformasikan. 2014;9:42-7. doi:10.15265/IY-2014-0018
30. Sandhu E, Weinstein S, McKethan A, dkk. Penggunaan sekunder data rekam kesehatan elektronik: Manfaat dan hambatan. Jt Comm J Qual Pasien Saf. 2012;38:34-40. 1
31. Schilling LM, Kwan BM, Drolshagen CT, dkk. Arsitektur yang dapat diskalakan untuk infrastruktur teknologi jaringan pertanyaan translasi federasi (SAFTINet) untuk jaringan data terdistribusi. EGEMS. 2013;1:1027. doi:10.13063/2327-9214.1027.
32. Kemitraan Hasil Medis Observasi. CDM dan Kosakata Standar (Versi 4). 2012. http://omop.org/VocabV4.5. Diakses 1 April 2013.
33. Pencocokan Skema dan Pemetaan. http://www.springer.com/computer/database+management+%26+information+retrieval/book/978-3-642-16517-7. Diakses 25 Mei 2014.
34. Fagin R. Pembalikan Pemetaan Skema. Dalam: Prosiding simposium ACM SIGMOD-SIGACT-SIGART kedua puluh lima tentang prinsip-prinsip sistem database. New York: ACM; 2006. hal. 50-9. doi: 10.1145/114351.1142359.
35. Häyrinen K, Saranto K, Nykänen P. Definisi, struktur, isi, penggunaan dan dampak catatan kesehatan elektronik: tinjauan literatur penelitian. Int J Med Inf. 2008;77:291-304. doi:10.1016/j.ijmedinf.2007.09.001.
36. Madhavan J, Jeffery SR, Cohen S, dkk. Integrasi data skala web: Anda hanya mampu membayar sesuai pemakaian. Dalam: Dalam Prok. dari CIDR-07; 2007.
37. Davidson SB, Overton C, Buneman P. Tantangan dalam mengintegrasikan sumber data biologis. J Comput Biol. 1995;2:557-72. doi:10.1089/cmb.1995.2.557.
38. Taniar D, Chen L, editor. Integrasi Data Warehousing, Data Mining dan Teknologi Database: Pendekatan Inovatif. IGI Global 2011. http://www.igi-global.com/chapter/survey-extract-transform-load-technology/53076. Diakses 24 Sep 2014.
39. Stang PE, Ryan PB, Racoosin JA, dkk. Memajukan ilmu untuk pengawasan aktif: alasan dan desain untuk Kemitraan Hasil Medis Observasi. Ann Intern Med. 2010;153:600-6. doi:10.1059/0003-4819-153-9-201011020-00010.
40. Rothstein MA. Arus dalam etika kontemporer. Penelitian privasi di bawah HIPAA dan aturan umum. J Hukum Med Etika. 2005;33:154-9.
41. Nass SJ, Levit LA, Gostin LO, dkk. Di luar aturan privasi HIPAA: meningkatkan privasi, meningkatkan kesehatan melalui penelitian. Washington, DC: Pers Akademi Nasional; 2009. https://www.nap.edu/read/12458/chapter/1
42. Ness RB. Pengaruh aturan privasi HIPAA pada penelitian kesehatan. JAMA. 2007;298:2164-70. doi:10.1001/jama.298.18.2164.
43. Herdman R, Moses HL, Amerika Serikat, dkk. Pengaruh aturan privasi HIPAA pada penelitian kesehatan: Prosiding lokakarya yang dipresentasikan ke Forum Kebijakan Kanker Nasional. Washington DC: Pers Akademi Nasional; 2006.
44. Selker HP, Pienta KJ. Pentingnya perubahan yang diusulkan dalam'Aturan Umum' untuk peneliti klinis dan translasi. Clin Transl Sci. 2011;4:312-3. doi:10.1111/j.1752-8062.2011.00352.x.
45. Wyatt L, Caufield B, Pol D. Prinsip untuk Tolok Ukur ETL. Dalam: Nambiar R, Poess M, eds. Evaluasi Kinerja dan Benchmarking. Springer Berlin Heidelberg 2009. 183-98.https://link.springer.com/book/10.1007%2F978-3642-10424-4. Diakses 1 Des 2014.
46. Kusen MR, Kwan BM, Menguap BP, dkk. Karakteristik rumah medis dan kontrol asma: protokol studi kohort observasional prospektif. Cuci EGEMS DC. 2013;1:1032. doi:10.13063/2327-9214.1032.
47. Kwan BM, Sills MR, Graham D, dkk. Keterlibatan Pemangku Kepentingan dalam Implementasi Ukuran PatientReported Outcomes (PRO): Laporan dari Jaringan Penelitian Berbasis Praktik SAFTINet (PBRN). J Am Board Fam Med JABFM. 2016;29:102-15. doi:10.3122/jabfm.2016.01.150141.
48. Kahn MG, Raebel MA, Glanz JM, dkk. Kerangka kerja pragmatis untuk penilaian kualitas data situs tunggal dan multi situs dalam penelitian klinis berbasis catatan kesehatan elektronik. Perawatan Medis. 2012;50(Pemasok):S21-9. doi:10.1097/MLR. 0b013e318257dd67.
49. Legris P, Ingham J, Collerette P. Mengapa orang menggunakan teknologi informasi? Sebuah tinjauan kritis dari model penerimaan teknologi. Manajemen Inf. 2003;40:191-204. doi:10.1016/S0378-7206(01)00143-4.
50. Fagin R, Haas LM, Hernández M, dkk. Clio: Pembuatan Pemetaan Skema dan Pertukaran Data. Dalam: Borgida AT, Chaudhri VK, Giorgini P, dkk., eds. Pemodelan Konseptual: Fondasi dan Aplikasi. Springer Berlin Heidelberg 2009. 198-236.http://link.springer.com/chapter/10.1007/978-3642-02463-4_12. Diakses 11 Juni 2014.