

OHDSI 2019 US Symposium オープニングセッション  
George Hripcsak 講演部分の前半より  
2019/9/15, North Bethesda, MD, USA.

(2019/10/27 版)



## Welcome to OHDSI 2019: This is our community

**George Hripcsak**, MD, MS, Chair of the Department of Biomedical Informatics at  
Columbia University Medical Center

**Harlan M. Krumholz**, MD, SM, Harold H. Hines, Jr. Professor of Medicine at Yale  
University School of Medicine; Director, Yale New Haven Hospital Center for  
Outcomes Research and Evaluation (CORE)  
@hmkyale

OHDSI is  
an open science community



OHDSI's mission

To improve health by empowering a  
community to collaboratively  
generate the evidence that promotes  
better health decisions and better care



## OHDSI's values

- **Innovation:** Observational research is a field which will benefit greatly from disruptive thinking. We actively seek and encourage fresh methodological approaches in our work.
  - **Reproducibility:** Accurate, reproducible, and well-calibrated evidence is necessary for health improvement.
  - **Community:** Everyone is welcome to actively participate in OHDSI, whether you are a patient, a health professional, a researcher, or someone who simply believes in our cause.
  - **Collaboration:** We work collectively to prioritize and address the real world needs of our community's participants.
  - **Openness:** We strive to make all our community's proceeds open and publicly accessible, including the methods, tools and the evidence that we generate.
  - **Beneficence:** We seek to protect the rights of individuals and organizations within our community at all times.
- 

### OHDSI 2019 へようこそ

(訳注:OHDSI は古代ギリシャ叙事詩 *Odyssey* にちなんでおり日本語読みはオデッセイ、英語発音は人/場面により異なりますが、オデシーからオウデシー、オウデッシーといったところです。)

OHDSI が初めての方へ、OHDSI はオープンサイエンスコミュニティです。

私たちの使命は、読み上げると、コミュニティがより良い健康判断とより良いケアを促進するエビデンスを共同で生成できるようにすることで、健康を改善することです。これは本当に私たちが行っていることであり、私はこのミッションステートメントにどのように到達したのか知らないのですが、私は実際それがかなり気に入っています。これは本当に私たちがそうしようとしていることを実際に要約しており、私たちはそれを維持することができ、願わくば長い間維持していきたい。このスライドは私達の価値を示していて読み上げませんが、イノベーション、再現性、コミュニティ、コラボレーション、オープン性、有益性があります。



## OHDSI community

We're all in this journey together...



256 collaborators in 27 different countries over six continents

ここに私たちのコミュニティの最新のスライドがあります。ここには、**27** か国に **256** 人のコラボレーターがいます。私はコラボレーターを、略歴を作成し、ウェブサイトに写真を載せてくれる人として定義しています。

27 カ国 6 大陸にいて、北米、ヨーロッパ、アジアに多く、アフリカ、南アメリカ、オーストラリアでも取り組んでいます。昨年申し上げたように、私たちは短期間、7 大陸で活動していました。しかし、南極大陸で取り組むと言っていた人が南極大陸を去ったので、私たちは 6 大陸にもどりました。来年はまた 7 大陸と言えるといいですね。



## OHDSI's community engagement

- Active community online discussion: [forums.ohdsi.org](https://forums.ohdsi.org)
  - >2,770 distinct users have made >18,700 posts on >3,250 topics
  - Implementers, Developers, Researchers, CDM Builders, Vocabulary users, OHDSI in Korea, OHDSI in China, OHDSI in Europe
- Weekly community web conferences for all collaborators to share their research ideas and progress
- >25 workgroups for solving shared problems of interest
  - ex: Common Data Model, Population-level Estimation, Patient-level Prediction, Phenotype, NLP, GIS, Oncology, Women of OHDSI
- Quarterly tutorials in OHDSI tools and best practices, taught by OHDSI collaborators for OHDSI collaborators
- OHDSI Symposiums held annually in North America, Europe and Asia to provide the community face-to-face opportunities to showcase research collaborations
- Follow us on Twitter @OHDSI and LinkedIn

私たちの取り組みについて少しお話したいと思います。私たちのフォーラムでは 3000 人近くの人が投稿していますので、コラボレーターは 256 人と述べましたがおそらく過小評価です。私たちは略歴数で示されるよりもるかに広いことが、3000 のトピックに関する 18,000 の投稿で示されています。各人が開発者や研究者などとして実装に関わりました。毎週のコミュニティ Web 会議がありますが、オデッセイの規模が大きいため、多くの場合、2 つの半球をカバーするために週に 2 回 Web 会議をする必要があります。実際に地球の周りを一周します。

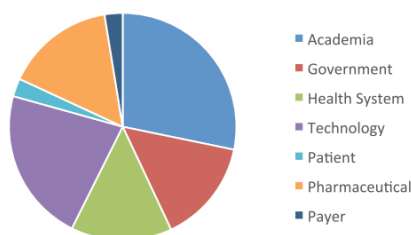
25 個以上のワーキンググループがあり、共通データモデル、人口レベル推定、患者レベル予測などがあります。

私たちは、オデッセイの共同研究者によって教えられる四半期ごとのチュートリアル、そして本気のツールとベストプラクティスを、一般的にボランティアベースで開催しています。まさにオープンサイエンスコミュニティとなっていて、多くのボランティアの努力とこれらのチュートリアルを履修した聴衆に感謝しています。そしてもちろん、毎年、北米、ヨーロッパ、アジアでシンポジウムを開催しています。そして、私たちはソーシャルメディア上で非常に強力な存在感を持っています。今では、Craig とのセッションに参加することで、常にやろうとしています。私たちはその存在感、コミュニティ、フォーラムへの投稿やアクティビティを毎年増やし続けています。

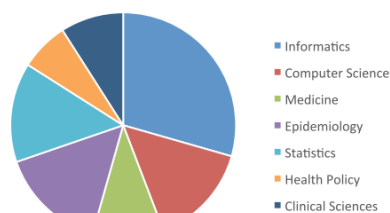


## Diversity of the OHDSI community represented today at the OHDSI Symposium

Stakeholder group



Disciplinary perspective



Relationship with OHDSI community	Persons
I am new to OHDSI and curious to learn more	240
I actively participate in OHDSI meetings and work groups	177
I use OHDSI tools and methods to support my research	176
I have an OMOP CDM instance	125
I am in the process of converting my data into the OMOP CDM	95
I actively participate in discussions on the OHDSI forum	74
I am participating in an OHDSI network research study	55
I contribute code to the OHDSI GitHub	48

今日のシンポジウムに登録されているのは誰なのか、左の円グラフに学術、政府、産業界からは医療システム、テクノロジー、医薬品、そして患者と保険者がいます。良いニュースは、学術、政府、産業が、それぞれ大きなスライスになっていることです。言い換えれば、医学でよく言われるように、3本の脚がある椅子になります。右側の円グラフでは、我々は情報

学、コンピューターサイエンス、医学、疫学などによる学際的なイニシアチブであることを示しています。

今日の会議の驚きのひとは、約 540 人の登録者のうちほぼ半数、約 240 人が OHDSI 初心者でもっと知りたいという人だということです。そして、残りの半分ほどは OHDSI 経験者でどのようなことをしてきたかが説明されています。コミュニティは非常に活発で、私たちはみな非常に感謝しています。



## Data across the OHDSI community

- 152 entries on [2019 OHDSI data network inventory](#)
- 133 different databases with patient-level data from various perspectives:
  - Electronic health records, administrative claims, hospital systems, clinical registries, health surveys, biobanks
- Data in 18 different countries, with >369 million patient records from outside US

**All using one open community data standard:  
OMOP Common Data Model**

---

私たちには国際的なデータネットワークがあります。それについて少し話させてください。OHDSI データネットワークリストには 152 のエントリがあります。フォーラムに 3000 人、我々の Web サイトに 250 人の略歴があるのと同様に、これらは我々の Web サイトで特定できるデータベースとなります。OHDSI-CDM または OMOP-CDM を使用するデータベースの総数はおそらくはるかに多くなりますが、それでもこのリストだけで患者レベルのデータを持つ 133 の異なるデータベースがあります。主に EHR と請求データですが、それだけでなく、臨床レジストリ、健康調査、バイオバンクもあり、データが増えています。

18 か国からのデータがあります。そして米国の外を見るならば。3.7 億件の患者記録があります。患者のほとんどは一意であり、各国で最大のデータベースをひとつずつだけ選んだ場合でも、各国間で重複はほとんどないでしょうから少なくとも 3 億人にはなります。他のデータベースの多くも重複していないため、それも過小評価されており、つまりもっと多くの米国内外のデータがあります。米国の各データは 300 万人から 2000 万人ですが、ほとんどが単一の人ではありません。これは合計数十億を反映する記録ですが、重複があるため、おそらく世界では 6 億人以上のユニークな患者を抱えています。世界は 70 億人です。ですから OHDSI 連携ネットワークが世界人口の 10%になるというポイントに非常に近づ

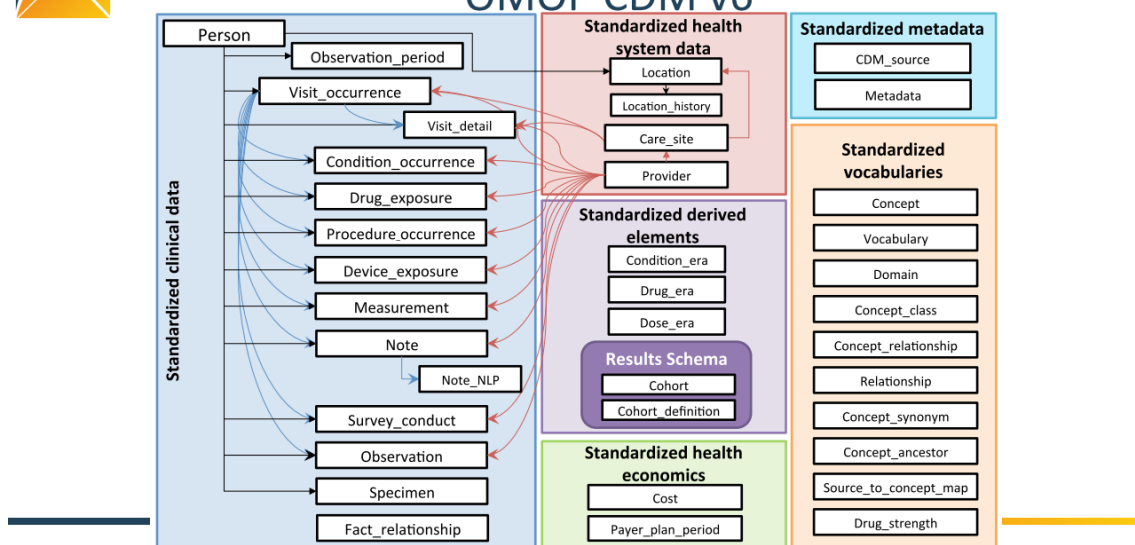
いています。これはすべて自発的なものです。それでも、データベースに存在する世界人口のおおよそ 10%を達成することができました。かなり良い状況です。

共通データモデルを用いて、世界中のデータベースセットにより世界の人口を獲得することも 1 つありますが、それだけでなく同一の語彙を使用してエレガントで単一のデータモデルでそれを行うことも私達の非常な成果です。



## Open community data standard:

### OMOP CDM v6



OHDSI/OMOP の一般的なデータモデルスキーマの画像を次に示します。その美しさはシンプルさにあることがわかります。理解しやすいです。

エンティティ属性値は確かにあります、他の場所に収まらないものはすべて入れるようなテーブルですが、それ以外は物事を分割して、OHDSI を初めて使用する研究者でもすべての状況を調べて、よしここは薬物だ、ここは疾病だ、デバイスだ、処置だ、使ったものだということができるようにしています。これは私たちの成功のひとつの要因だと思います。

(訳注：OHDSI のデータモデルはかつて OMOP initiative にて策定されたものを OHDSI が引き継いで発展させたものです。由来から今でも OMOP Common Data Model と呼んでいます。OMOP の英語発音はオモップからオウモップ。)





## OHDSI's standardized vocabularies

- >130 Vocabularies across 40 domains
  - MU3 standards: SNOMED, RxNorm, LOINC
  - Disparate sources: ICD9CM, ICD10(CM), Read, NDC, Gemscript, CPT4, HCPCS...
- >7.4 million concepts
  - >3.0 million standard concepts
  - >3.8 million source codes
  - >511,000 classification concepts
- >45 million concept relationships
- >74 million ancestral relationships

---

Publicly available for download at: <http://athena.ohdsi.org/>

世界のデータベースを作成するには世界の語彙をサポートする必要があり、40 のドメインで 130 の語彙をサポートしています。それにより、世界中のすべてのものが、3 つの SNOMED、RxNorm、LOINC と他のボキャブラリーのセットを使用する標準セットにマッピングされます。そうして世界中で広く採用されている共通のセットに入ります。

700 万の概念のうち約半数は標準語彙で、半数がそれら標準にマップされる語彙です。強調したいのは 50 万の分類概念で、知識工学は OHDSI に非常に重要なのですが、これは 700 万を整理するために使用される知識概念のような 50 万です。

規模が非常に大きく、語彙に貢献する OHDSI メンバーに感謝します。彼らは 4500 万の概念関係、7400 万の祖先関係からなるこの大きなものを整理しています。その多くは語彙間のマッピングに入れなければなりません。

(訳注 : OHDSI で言う語彙 Vocabulary とは、日本でいえばレセプト電算コードに代表される、コードと名称/用語/概念の対応集合のことです。用語集にコードをつけたものとも言えます。)



## Highlights of progress from the community: Data standards

- Increased adoption of OMOP CDM
- Evaluation of vocabulary
- Expanded vocabulary
- Community collaboration around conventions (THEMIS)
- Added rigor around data quality (see Clair and Andrew)

---

昨年に起こったことのいくつかのハイライトを見てみましょう。私たちは、共通データモデル評価の採用の増加、語彙の評価、語彙の拡張を行いました。

とびきりのグループが約束事に取り組んでいます。標準に取り組んでいる人は知っていると思いますが、標準があって、2人の人が彼らのデータベースを同じ標準に変換する場合、たとえ同じ施設の人だとしても、その2人はほぼ何も一致しません。そこが約束事の出番で、データとモデルを与えられたとき、片方をもう片方にどうマッピングするのか、どんな判断をするのか、そして最も一般的な判断から始めることで、非常に大きく非常に多様なネットワークにわたって実際に一貫性を持たせることができます。

それから、データ品質に関する厳格さを追加しました。これについては、最初の基調講演でご紹介します。





## Research and Applications

### Effect of vocabulary mapping for conditions on phenotype cohorts

George Hripcsak<sup>1,2,3</sup> Matthew E Levine<sup>1,2</sup> Ning Shang<sup>1,2</sup> and Patrick B Ryan<sup>1,2,4</sup>

<sup>1</sup>Department of Biomedical Informatics, Columbia University, New York, New York, USA, <sup>2</sup>Observational Health Data Sciences and Informatics (OHDSI), New York, New York, USA, <sup>3</sup>Medical Informatics Services, NewYork-Presbyterian Hospital, New York, New York, USA, and <sup>4</sup>Epidemiology Analytics, Janssen Research and Development, Titusville, New Jersey, USA

Corresponding Author: George Hripcsak, MD, MS, Department of Biomedical Informatics, Columbia University Irving Medical Center, 622 W 168th St, PH20, New York, NY 10032, USA (hripcsak@columbia.edu)

Received 27 April 2018; Revised 13 August 2018; Editorial Decision 22 August 2018; Accepted 3 September 2018

#### ABSTRACT

**Objective:** To study the effect on patient cohorts of mapping condition (diagnosis) codes from source billing vocabularies to a clinical vocabulary.

**Materials and Methods:** Nine International Classification of Diseases, Ninth Revision, Clinical Modification (ICD9-CM) concept sets were extracted from eMERGE network phenotypes, translated to Systematized Nomenclature of Medicine - Clinical Terms concept sets, and applied to patient data that were mapped from source ICD9-CM and ICD10-CM codes to Systematized Nomenclature of Medicine - Clinical Terms codes using Observational Health Data Sciences and Informatics (OHDSI) Observational Medical Outcomes Partnership (OMOP) vocabulary mappings. The original ICD9-CM concept set and a concept set extended to ICD10-CM were used to create patient cohorts that served as gold standards.

**Results:** Four phenotype concept sets were able to be translated to Systematized Nomenclature of Medicine - Clinical Terms without ambiguities and were able to perform perfectly with respect to the gold standards. The other 5 lost performance when 2 or more ICD9-CM or ICD10-CM codes mapped to the same Systematized Nomenclature of Medicine - Clinical Terms code. The patient cohorts had a total error (false positive and false negative) of up to 0.15% compared to querying ICD9-CM source data and up to 0.26% compared to querying ICD9-CM and ICD10-CM data. Knowledge engineering was required to produce that performance; simple automated methods to generate concept sets had errors up to 10% (one outlier at 280%).

**Discussion:** The translation of data from source vocabularies to Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) resulted in very small error rates that were an order of magnitude smaller than other error sources.

**Conclusion:** It appears possible to map diagnoses from disparate vocabularies to a single clinical vocabulary and carry out research using a single set of definitions, thus improving efficiency and transportability of research.

ここにお示しするのは語彙マッピングに関する論文であり、コホートの作成に有効であることを示しています。



### HemOnc: A new standard vocabulary for chemotherapy regimen representation in the OMOP common data model

Jeremy L. Warner<sup>a,b,\*</sup>, Dmitry Dymshyts<sup>c</sup>, Christian G. Reich<sup>d</sup>, Michael J. Gurley<sup>e</sup>, Harry Hochheiser<sup>f</sup>, Zachary H. Moldwin<sup>g</sup>, Rimma Belenkaya<sup>h</sup>, Andrew E. Williams<sup>i</sup>, Peter C. Yang<sup>b,j</sup>

<sup>a</sup> Vanderbilt University Medical Center, Nashville, TN, United States

<sup>b</sup> HemOnc.org, LLC, Lexington, MA, United States

<sup>c</sup> Odyssey Data Services, Inc., Cambridge, MA, United States

<sup>d</sup> IQVIA, Cambridge, MA, United States

<sup>e</sup> Northwestern University, Chicago, IL, United States

<sup>f</sup> University of Pittsburgh, Pittsburgh, PA, United States

<sup>g</sup> University of Illinois at Chicago College of Pharmacy, Chicago, IL, United States

<sup>h</sup> Memorial Sloan Kettering Cancer Center, New York, NY, United States

<sup>i</sup> Tufts University, Medford, MA, United States

<sup>j</sup> Massachusetts General Hospital, Harvard Medical School, Boston, MA, United States



これは OMOP の共通データモデルからの化学療法レジメンの表現に関する論文です。私たちの腫瘍学 WG は、この 2 年間非常に活発に活動しています。




## Highlights of progress from the community: Methods research

- Phenotype definition
- Phenotype evaluation
- Study design evaluation

手法研究として、phenotyping の開発と phenotyping の評価と研究デザイン評価のいくつかの例をご紹介します。






Contents lists available at ScienceDirect

Journal of Biomedical Informatics


journal homepage: [www.elsevier.com/locate/jbin](http://www.elsevier.com/locate/jbin)



### Facilitating phenotype transfer using a common data model

George Hripesak<sup>a,b,\*</sup>, Ning Shang<sup>a</sup>, Peggy L. Peissig<sup>c</sup>, Luke V. Rasmussen<sup>d</sup>, Cong Liu<sup>a</sup>, Barbara Benoit<sup>e</sup>, Robert J. Carroll<sup>f</sup>, David S. Carrell<sup>g</sup>, Joshua C. Denny<sup>h,i</sup>, Ozan Dikilitas<sup>j</sup>, Vivian S. Gainer<sup>k</sup>, Kayla Marie Howell<sup>l</sup>, Jeffrey G. Klann<sup>e</sup>, Iftikhar J. Kullo<sup>l</sup>, Todd Lingren<sup>k</sup>, Frank D. Mentch<sup>j</sup>, Shawn N. Murphy<sup>g</sup>, Karthik Natarajan<sup>a,b</sup>, Jennifer A. Pacheco<sup>d</sup>, Wei-Qi Wei<sup>l</sup>, Ken Wiley<sup>m</sup>, Chunhua Weng<sup>a</sup>

<sup>a</sup>Department of Biomedical Informatics, Columbia University, New York, NY, United States  
<sup>b</sup>Medical Informatics Services, NewYork-Presbyterian Hospital, New York, NY, United States  
<sup>c</sup>Center for Precision Medicine Research, Marshfield Clinic Research Institute, Marshfield, WI, United States  
<sup>d</sup>Northwestern University Feinberg School of Medicine, Chicago, IL, United States  
<sup>e</sup>Research Information Science and Computing, Partners Healthcare, Boston, MA, United States  
<sup>f</sup>Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States  
<sup>g</sup>Kaiser Permanente Washington Health Research Institute, Seattle, WA, United States  
<sup>h</sup>Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, United States  
<sup>i</sup>Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN, United States  
<sup>j</sup>Vanderbilt Institute for Clinical and Translational Research, Vanderbilt University Medical Center, Nashville, TN, United States  
<sup>k</sup>Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States  
<sup>l</sup>Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA, United States  
<sup>m</sup>National Human Genome Research Institute, NIH, Bethesda, MD, United States



米国の eMERGE ネットワークと提携して phenotyping を研究しています。



Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: [www.elsevier.com/locate/yjbin](http://www.elsevier.com/locate/yjbin)



## PheValuator: Development and evaluation of a phenotype algorithm evaluator

Joel N. Swerdel<sup>a,b,\*</sup>, George Hripcsak<sup>b,c</sup>, Patrick B. Ryan<sup>a,b,c</sup>

<sup>a</sup>Janssen Research & Development, 920 Route 202, Raritan, NJ 08869, USA

<sup>b</sup>OHDSI Collaborators, Observational Health Data Sciences and Informatics (OHDSI), 622 West 168th Street, PH-20, New York, NY 10032, USA

<sup>c</sup>Columbia University, 622 West 168th Street, PH20, New York, NY 10032, USA

### ARTICLE INFO

**Keywords:**  
Phenotype algorithms  
Validation  
Diagnostic predictive modeling

### ABSTRACT

**Background:** The primary approach for defining disease in observational healthcare databases is to construct phenotype algorithms (PAs), rule-based heuristics predicated on the presence, absence, and temporal logic of clinical observations. However, a complete evaluation of PAs, i.e., determining sensitivity, specificity, and positive predictive value (PPV), is rarely performed. In this study, we propose a tool (PheValuator) to efficiently estimate a complete PA evaluation.

**Methods:** We used 4 administrative claims datasets: OptumInsight's de-identified Clinformatics™ Datamart (Eden Prairie, MN); IBM MarketScan Multi-State Medicaid; IBM MarketScan Medicare Supplemental Beneficiaries; and IBM MarketScan Commercial Claims and Encounters from 2000 to 2017. Using PheValuator involves (1) creating a diagnostic predictive model for the phenotype, (2) applying the model to a large set of randomly selected subjects, and (3) comparing each subject's predicted probability for the phenotype to inclusion/exclusion in PAs. We used the predictions as a 'probabilistic gold standard' measure to classify positive/negative cases. We examined 4 phenotypes: myocardial infarction, cerebral infarction, chronic kidney disease, and atrial fibrillation. We examined several PAs for each phenotype including 1-time (1X) occurrence of the diagnosis code in the subject's record and 1-time occurrence of the diagnosis in an inpatient setting with the diagnosis code as the primary reason for admission (1X-IP-1stPos).

**Results:** Across phenotypes, the 1X PA showed the highest sensitivity/lowest PPV among all PAs. 1X-IP-1stPos yielded the highest PPV/lowest sensitivity. Specificity was very high across algorithms. We found similar results between algorithms across datasets.

**Conclusion:** PheValuator appears to show promise as a tool to estimate PA performance characteristics.

こちらは、Phenotyping の評価の手間の削減。少し後でお話しますが、結局のところ、Phenotyping やコホート定義を評価するには手作業でゴールドスタンダードを作るためものすごく手間がかかります。どうやってその手間を削減できるのでしょうか。



Received: 2 August 2018 | Revised: 29 April 2019 | Accepted: 3 May 2019  
DOI: 10.1002/smi.8213

### RESEARCH ARTICLE

WILEY Statistics  
in Medicine

## A plea to stop using the case-control design in retrospective database studies

Martijn J. Schuemie<sup>1,2,3</sup> | Patrick B. Ryan<sup>1,2,4</sup> | Kenneth K.C. Man<sup>5,6,7,8</sup> | Ian C.K. Wong<sup>5,6</sup> | Marc A. Suchard<sup>1,3,9,10</sup> | George Hripcsak<sup>1,4,11</sup>

<sup>1</sup>Observational Health Data Sciences and Informatics, New York, New York

<sup>2</sup>Epidemiology Analytics, Janssen Research and Development, Titusville, New Jersey

<sup>3</sup>Department of Biostatistics, University of California, Los Angeles, California

<sup>4</sup>Department of Biomedical Informatics, Columbia University Medical Center, New York, New York

<sup>5</sup>Centre for Safe Medication Practice and Research, Department of Pharmacology and Pharmacy, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Pokfulam, Hong Kong

<sup>6</sup>Research Department of Practice and Policy, UCL School of Pharmacy, London, UK

<sup>7</sup>Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands

<sup>8</sup>Department of Social Work and Social Administration, Faculty of Social Sciences, The University of Hong Kong, Pokfulam, Hong Kong

<sup>9</sup>Department of Biostatistics, University of California, Los Angeles, California

The case-control design is widely used in retrospective database studies, often leading to spectacular findings. However, results of these studies often cannot be replicated, and the advantage of this design over others is questionable. To demonstrate the shortcomings of applications of this design, we replicate two published case-control studies. The first investigates isotretinoin and ulcerative colitis using a simple case-control design. The second focuses on dipeptidyl peptidase-4 inhibitors and acute pancreatitis, using a nested case-control design. We include large sets of negative control exposures (where the true odds ratio is believed to be 1) in both studies. Both replication studies produce effect size estimates consistent with the original studies, but also generate estimates for the negative control exposures showing substantial residual bias. In contrast, applying a self-controlled design to answer the same questions using the same data reveals far less bias. Although the case-control design in general is not at fault, its application in retrospective database studies, where all exposure and covariate data for the entire cohort are available, is unnecessary, as other alternatives such as cohort and self-controlled designs are available. Moreover, by focusing on cases and controls it opens the door to inappropriate comparisons between exposure groups, leading to confounding for which the design has few options to adjust for. We argue that this design should no longer be used in these types of data. At the very least, negative control exposures should be used to prove that the concerns raised here do not apply.

OHDSI コミュニティから他のコミュニティに手を差し伸べるもの。例えば、ケースコントロールデザインの誤用。他のコンテキストでは有用ですが、レトロスペクティブなデータベース研究では使い出がありません。一般的にはケースコントロールスタディを使用する理由はなく、バイアスと検出力の損失につながります。



## Highlights of progress from the community: Open source development

- ATLAS 2.7.3 released
- Criteria2Query published
- Community contributions for multiple OMOP CDM utilities

そして、オープンソースソフトウェアの開発。我々の大きなツール **Atlas** は、バージョン 2.7.3 がリリースされました。ほかのツール、**Criteria2Query** が公開されました。そして、OMOP-CDM を使用するユーティリティに対するコミュニティのさまざまな貢献が沢山あります。



← → ↻ atlas.ohdsi.org/#/home ☆ 🔍 📱 🌐 🌐 🌐 🌐 🌐

**ATLAS** Home | ryan@ohdsi.org

Home

Welcome to ATLAS.  
ATLAS is an open source application developed as a part of [OHDSI](#) intended to provide a unified interface to patient level data and analytics.

Documentation  
The ATLAS user guide can be found [here](#).

Getting Started

Define a New Cohort Begin performing research by defining the group of people you intend to study

Search the Vocabulary Search the different ontologies used to describe patient level data around the world

Release Notes

[ATLAS Version 2.7.3 Release Notes](#)  
[WebAPI Version 2.7.3 Release Notes](#)

This latest release contains 7 feature enhancements and issue resolutions:

- Cohort definitions creation date is 4 hours greater than actual while being on EST timezone
- Do not call user/refresh endpoint case of IAP authentication
- Characterization pop-up shows wrong percentage
- Role import / export works incorrectly
- Title Consistency
- Active Directory groups mapping issue
- Cannot save concept set modification in cohort definition

Apache 2.0  
open source software  
provided by  
**OHDSI**  
join the journey



## Research and Applications

### Criteria2Query: a natural language interface to clinical databases for cohort definition

Chi Yuan,<sup>1,2</sup> Patrick B. Ryan,<sup>1,2</sup> Casey Ta,<sup>1</sup> Yixuan Guo,<sup>1</sup> Ziran Li,<sup>1</sup> Jill Hardin,<sup>3</sup>  
Rupa Makadia,<sup>2</sup> Peng Jin,<sup>1</sup> Ning Shang,<sup>1</sup> Tian Kang,<sup>1</sup> and Chunhua Weng<sup>1</sup>

<sup>1</sup>Department of Biomedical Informatics, Columbia University, New York, New York, USA, <sup>2</sup>Department of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing, Jiangsu Province, P.R. China, and <sup>3</sup>Epidemiology Analytics, Janssen Research and Development, Titusville, New Jersey, USA

Corresponding Author: Chunhua Weng, PhD, Department of Biomedical Informatics, Columbia University, 622 West 168th Street, PH-20, Room 407, New York, NY 10032, USA (chunhua@columbia.edu)

Received 7 September 2018; Revised 18 November 2018; Editorial Decision 27 November 2018; Accepted 29 November 2018

#### ABSTRACT

**Objective:** Cohort definition is a bottleneck for conducting clinical research and depends on subjective decisions by domain experts. Data-driven cohort definition is appealing but requires substantial knowledge of terminologies and clinical data models. Criteria2Query is a natural language interface that facilitates human-computer collaboration for cohort definition and execution using clinical databases.

**Materials and Methods:** Criteria2Query uses a hybrid information extraction pipeline combining machine learning and rule-based methods to systematically parse eligibility criteria text, transforms it first into a structured criteria representation and next into sharable and executable clinical data queries represented as SQL queries conforming to the OMOP Common Data Model. Users can interactively review, refine, and execute queries in the ATLAS web application. To test effectiveness, we evaluated 125 criteria across different disease domains from ClinicalTrials.gov and 52 user-entered criteria. We evaluated F1 score and accuracy against 2 domain experts and calculated the average computation time for fully automated query formulation. We conducted an anonymous survey evaluating usability.

**Results:** Criteria2Query achieved 0.795 and 0.895 F1 score for entity recognition and relation extraction, respectively. Accuracies for negation detection, logic detection, entity normalization, and attribute normalization were 0.984, 0.864, 0.514 and 0.793, respectively. Fully automatic query formulation took 1.22 seconds/criterion. More than 80% (11+ of 13) of users would use Criteria2Query in their future cohort definition tasks.

**Conclusions:** We contribute a novel natural language interface to clinical databases. It is open source and supports fully automated and interactive modes for autonomous data-driven cohort definition by researchers with minimal human effort. We demonstrate its promising user friendliness and usability.



## Data and text mining

### PatientExploreR: an extensible application for dynamic visualization of patient clinical history from electronic health records in the OMOP common data model

Benjamin S. Glicksberg<sup>1</sup>, Boris Oskotsky<sup>1</sup>, Phyllis M. Thangaraj<sup>2,3,4,1</sup>,  
Nicholas Giangreco<sup>2,3,4,1</sup>, Marcus A. Badgeley<sup>5,7</sup>,  
Kipp W. Johnson<sup>5,1</sup>, Debajyoti Datta<sup>1</sup>, Vivek A. Rudrapatna<sup>1,6</sup>,  
Nadav Rappoport<sup>1</sup>, Mark M. Shervey<sup>5</sup>, Riccardo Miotto<sup>5</sup>,  
Theodore C. Goldstein<sup>1</sup>, Eugenia Rutenberg<sup>1</sup>, Remi Frazier<sup>7</sup>,  
Nelson Lee<sup>2</sup>, Sharat Israni<sup>1</sup>, Rick Larsen<sup>2</sup>, Bethany Percha<sup>5</sup>, Li Li<sup>5</sup>,  
Joel T. Dudley<sup>5</sup>, Nicholas P. Tatonetti<sup>2,3,4</sup> and Atul J. Butte<sup>1,8,\*</sup>

<sup>1</sup>Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA 94158, USA, <sup>2</sup>Department of Biomedical Informatics, <sup>3</sup>Department of Systems Biology, <sup>4</sup>Department of Medicine, Columbia University, New York, NY 10032, USA, <sup>5</sup>Departments of Genomics and Data Science, Icahn Institute for Genomic Sciences and Multiscale Biology, Icahn School of Medicine at Mount Sinai, Institute of Next Generation Healthcare, New York, NY 10022, USA, <sup>6</sup>Division of Gastroenterology, Department of Medicine, University of California, San Francisco, CA 94158, USA, <sup>7</sup>Enterprise Information and Analytics, University of California, San Francisco, San Francisco, CA 94158, USA and <sup>8</sup>Center for Data-Driven Insights and Innovation, University of California Health, Oakland, CA 94607, USA

## Application Notes

### ROMOP: a light-weight R package for interfacing with OMOP-formatted electronic health record data

Benjamin S. Glicksberg,<sup>1</sup> Boris Oskotsky,<sup>1</sup> Nicholas Giangreco,<sup>2,1</sup>  
Phyllis M. Thangaraj,<sup>2,1</sup> Vivek Rudrapatna,<sup>1</sup> Debajyoti Datta,<sup>1</sup> Remi Frazier,<sup>3</sup>  
Nelson Lee,<sup>2</sup> Rick Larsen,<sup>2</sup> Nicholas P. Tatonetti<sup>2</sup> and Atul J. Butte<sup>1</sup>

<sup>1</sup>Department of Pediatrics Bakar Computational Health Sciences Institute, University of California San Francisco, San Francisco, California, USA, <sup>2</sup>Departments of Biomedical Informatics, Systems Biology, and Medicine, Columbia University, New York, New York, USA and <sup>3</sup>Academic Research Systems, Department of Enterprise Data Warehouse University of California San Francisco, San Francisco, California, USA

\*These two authors contributed equally to the study.

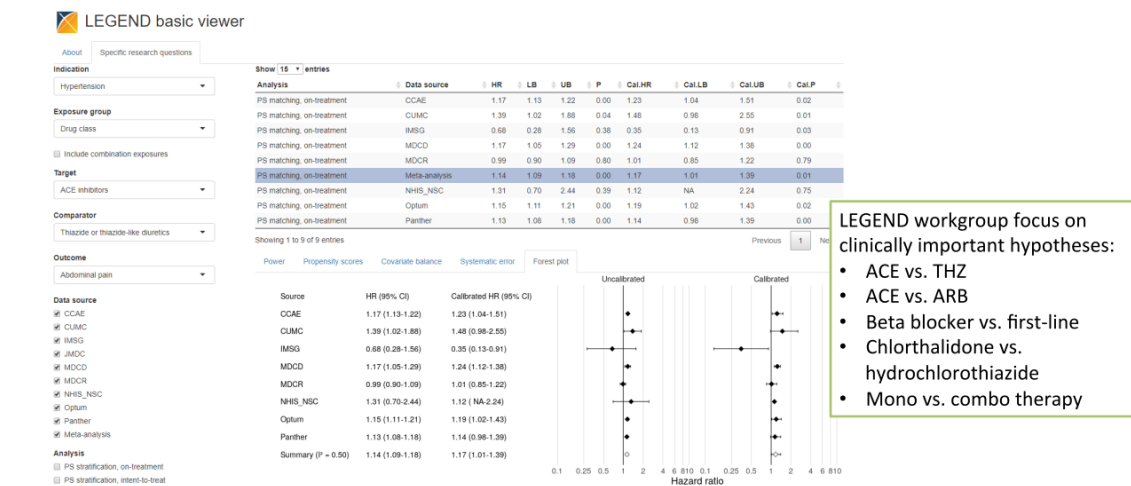
Corresponding Author: Atul J. Butte, MD, PhD, Bakar Computational Health Sciences Institute, University of California San Francisco, San Francisco, CA 94158, USA (Atul.Butte@ucsf.edu)

Received 3 July 2018; Revised 26 October 2018; Editorial Decision 29 November 2018; Accepted 2 December 2018

これは Atlas の新しいバージョンです。これは、医療情報学のトップジャーナルである JAMIA に掲載された Criteria2Query に関する論文です。そして、ツールのセットが、オープンソースコミュニティとして OMOP-CDM を活用して構築されたツールグループから出されていて、物事がどんどん成長しています。



## Highlights of progress from the community: Clinical applications



<http://data.ohdsi.org/LegendBasicViewer/>

臨床応用では、さきほど言ったように、実際にエビデンスを生成し、昨年はこのエビデンスの生成と報告にほとんどの時間を費やしました。スライドの右側の箱に色々ありますが、ACE 阻害剤とサイアザイドではサイアザイドが優れているようです。世界の高血圧症の人たちの半分は ACE 阻害剤で開始するのですが、おそらくそうすべきではないのです。これは大きな影響がある結果です。ACE 阻害剤と ARB では ARB が安全性で優れている可能性があります。ベータ遮断薬は第一選択薬の中核ではありませんが、研究者の主張とは違って賛美されるほどよくなく、単剤療法が良いかもしれません。LEGEND 高血圧研究から導かれたこの最初の非常に大規模な臨床研究論文はランセットにアクセプトされ、今後数週間で目にするができると思います。

Lancet 論文公開されました。(2019/10/24)

Lancet:

[https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(19\)32317-7/fulltext](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(19)32317-7/fulltext)

OHDSI Release:

<https://www.ohdsi.org/ohdsi-news-updates/legend-hypertension-study/>

(前半終わり)



後半のハイライトスライドをいくつかお示しします。



## FDA Biologics Effectiveness and Safety (BEST) Initiative

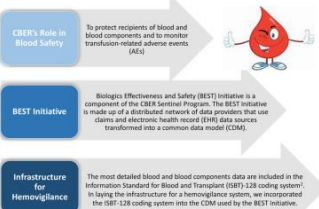
### Biologics Effectiveness and Safety (BEST) Initiative: Incorporating ISBT-128 Codes into OHDSI's OMOP Common Data Model to Build a National Hemovigilance System to Monitor Transfusion-Related Adverse Events

Jayce Obidi\*, Kinvera Chada\*, Joann Gruber\*, Graga Dore\*, Alan Williams\*, Emily Storch\*, Juan M Banda\*, Saurabh Gembur\*, Deepa Bahaj\*, Ross Hayden\*, Paul Bloudek\*, Shaan Granelli\*, George Hefcatal\*, Thomas Falcione\*, Karthik Natarajan\*, Dmitry Dymov\*, Sara Domagala\*, Christian Reiss\*, Nandini Sahani\*, Nerissa Williams\*, Steven Anderson\*, Azadeh Shoaibi\*

\*Center for Biologics Evaluation and Research, Food and Drug Administration, Silver Spring, MD, USA; \*Stanford University, Stanford, CA, USA; \*Regeneron Institute, Indianapolis, Indiana, USA; \*Columbia University, New York, NY, USA; \*Observational Health Data Sciences and Informatics, New York, NY, USA; \*Odyseus Data Services Inc., Cambridge, MA, USA; \*IQVIA, Cambridge, MA, USA

#### INTRODUCTION

The U.S. FDA Center for Biologics Evaluation and Research (CBER) regulates collection of whole blood and blood components utilized in transfusion<sup>1</sup>.

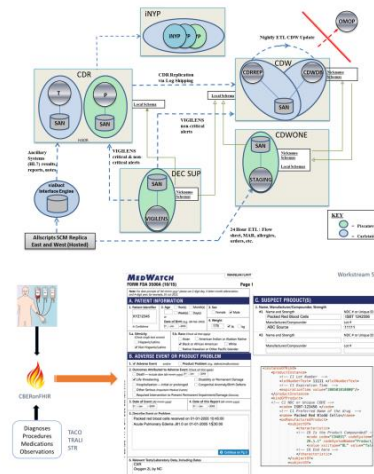


#### OBJECTIVE

The aim of this study was to build a component of the infrastructure for a national hemovigilance system using EHR data sources to monitor transfusion-related AEs by incorporating the ISBT-128 coding system into the Observational Medical Outcomes Partnership (OMOP) common data model (CDM) of the Observational Health Data Sciences and Informatics (OHDSI) consortium<sup>1</sup>.

#### METHODS

The CBER BEST Initiative is a collaboration with IQVIA, OHDSI Consortium, Columbia University, Stanford University, Indiana University, Regeneron Institute, Georgia Institute of Technology, and University of California Los Angeles. Within the BEST Initiative, we used three EHR databases that cover approximately 24 million patient records from geographically diverse areas of the U.S. We added a library of 14,543 ISBT-128 codes to the OMOP CDM. Each EHR data source requested access to its corresponding blood bank data and transformed its data into the OMOP CDM containing the newly added ISBT-128 codes. By querying the databases, we determined the type and frequency of ISBT-128 codes used in patient records from 2010-2017 within the blood banks of EHR data providers participating in the BEST Initiative.



## NIH All of Us Research Program

U.S. Department of Health & Human Services | National Institutes of Health



National Institutes of Health  
All of Us Research Program

ABOUT | FUNDING | NEWS, EVENTS, & MEDIA

JoinAllOfUs.org

Search

- 1,000,000 diverse participants
- Clinical data in OMOP CDM



### The future of health begins with you

The All of Us Research Program is a historic effort to gather data from one million or more people living in the United States to accelerate research and improve health. By taking into account individual differences in lifestyle, environment, and biology, researchers will uncover paths toward delivering precision medicine.

JOIN NOW



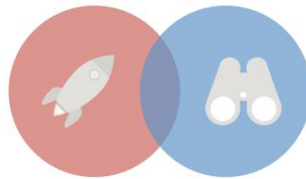


# The European Health Data and Evidence



## Mission

Our mission is to provide a new paradigm for the discovery and analysis of health data in Europe, by building a large-scale, federated network of data sources standardised to a common data model



## Vision

The European Health Data & Evidence Network (EHDEN) aspires to be the trusted observational research ecosystem to enable better health decisions, outcomes and care



## Objectives



### Harmonisation

Harmonise in excess of 100 million anonymised health records to the OMOP common data model, supported by an ecosystem of certified SMEs, and technical architecture for a federated network



### Evidence

Impact our understanding of, and improvement of, clinical outcomes for patients within diverse healthcare systems in the EU

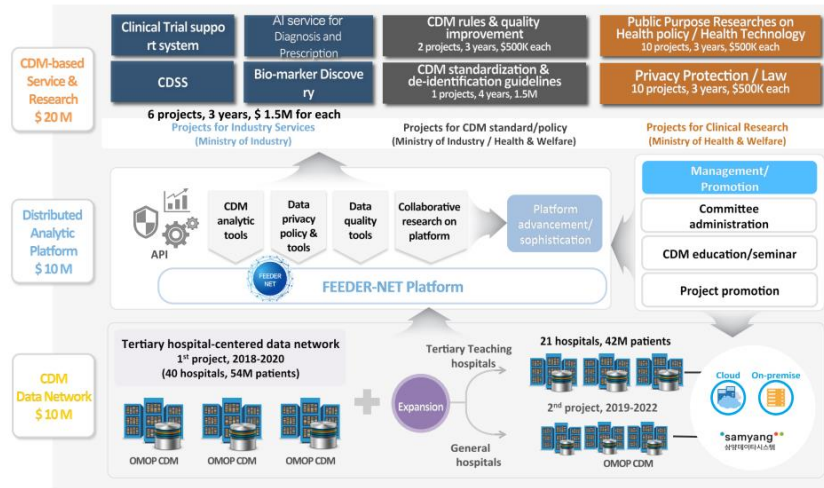


### Community

Establish a self-sustaining open science collaboration in Europe, supporting academia, industry, regulators, payers, government, NGOs and others



## National CDM Projects in Korea 2018-2022



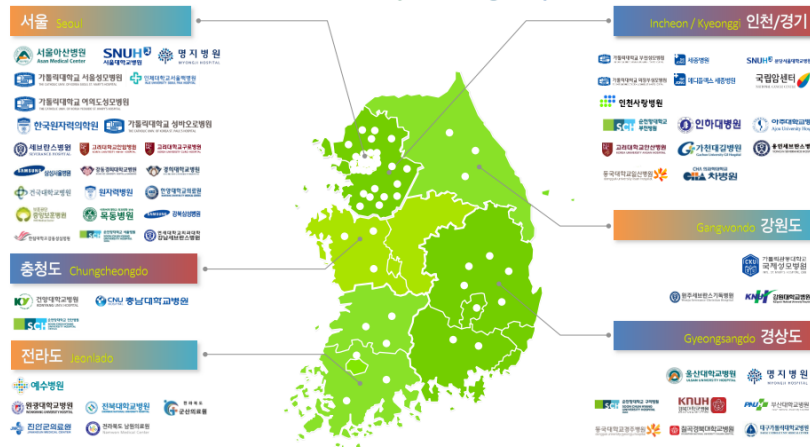
FEEDER-NET: Federated E-health big Data for Evidence Renovation NETWORK

37

## FEEDER-NET Data Network in Korea

Data Network of 60+ Hospitals, 98M Patients

70% of Tertiary Teaching Hospitals



38



## OHDSI China Symposium 2019



 **OHDSI**  
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

# European OHDSI Symposium 2020

27-29 March 2020 – Oxford, UK

Mathematical Institute,  
University of Oxford

 **NDORMS**  
NORFOLK DEPARTMENT OF ORTHOPAEDICS,  
TRUTH AND MUSCULOSKELETAL SCIENCES



## OHDSI evaluates itself and publishes the results

- OMOP CDM vocabulary evaluation
    - Automated translation of database works
    - Best not to automated the translation of cohort definitions
  - eMERGE phenotype implementation
    - Without CDM, narrative+flowchart+pseudocode+code list -> inconsistent
    - With CDM, can improve consistency and efficiency but caveats
  - PheValuator phenotype evaluation
    - Can estimate performance without manually curating gold standard
    - Estimates are imperfect
- 



## Key Challenges

- Data quality
  - Data spectrum
  - Causal inference
  - Communication/Education
  - Application
-





## What do I love about OHDSI?

- Spirit of collaboration, kindness, generosity



- Principles of transparency, open science, integrity



- Scientific rigor; reproducibility, validity

