

OHDSI RWE Revolution:

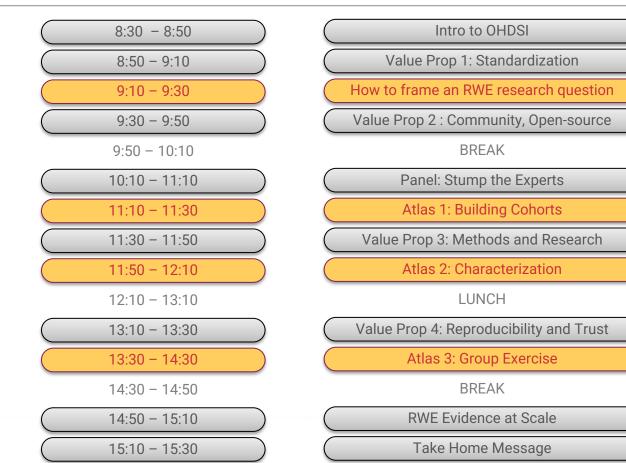
Igniting Data Modernization with Harmonized Standards

for Cutting-Edge Health Research

11-Nov-2023

Agenda







Introduction to OMOP and OHDSI

Or: What is it and why should you consider it?

11-Nov-2023





I disclose the following relevant relationship with commercial interests:

CEO of Odysseus Data Services





Observational Medical Outcomes Partnership

The OMOP Experiment





79 of 4,000 Vioxx users suffered heart problems or died

FDAAA calls for establishing Risk Identification and Analysis System

SEC. 905. ACTIVE POSTMARKET RISK IDENTIFICATION AND ANALYSIS.

(a) IN GENERAL.—Subsection (k) of section 505 of the Federal Food, Drug, and Cosmetic Act (21 U.S.C. 355) is amended by adding at the end the following:

"(3) ACTIVE POSTMARKET RISK IDENTIFICATION .---

"(A) DEFINITION.—In this paragraph, the term 'data' refers to information with respect to a drug approved under this section or under section 351 of the Public Health Service Act, including claims data, patient survey data, standardized analytic files that allow for the pooling and analysis of data from disparate data environments, and any other data deemed appropriate by the Secretary. "(B) DEVELOPMENT OF POSTMARKET RISK IDENTIFICA-

"(B) DEVELOPMENT OF POSTMARKET RISK IDENTIFICA-TION AND ANALYSIS METHODS.—The Secretary shall, not later than 2 years after the date of the enactment of the Food and Drug Administration Amendments Act of 2007, in collaboration with public, academic, and private entities—

"(i) develop methods to obtain access to disparate data sources including the data sources specified in subparagraph (C);

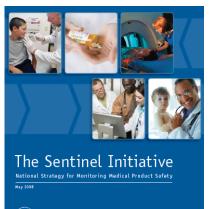
"(ii) develop validated methods for the establishment of a postmarket risk identification and analysis system to link and analyze safety data from multiple sources, with the goals of including, in aggregate—

"(I) at least 25,000,000 patients by July 1, 2010; and

"(II) at least 100,000,000 patients by July 1, 2012; and

"(iii) convene a committee of experts, including individuals who are recognized in the field of protecting data privacy and security, to make recommendations to the Secretary on the development of tools and methods for the ethical and scientific uses for, and communication of, postmarketing data specified under subparagraph (C), including recommendations on the development of effective research methods for the study of drug safety questions.

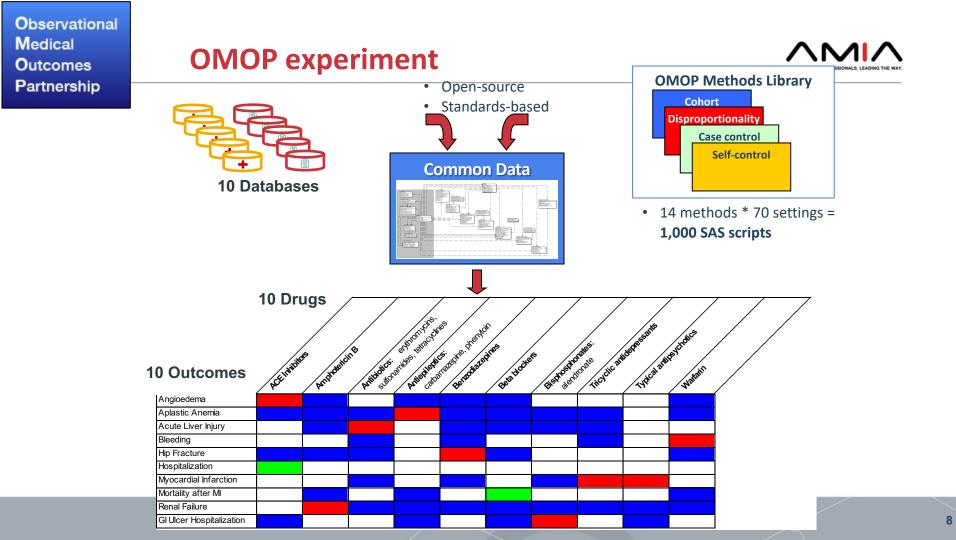
"(C) ESTABLISHMENT OF THE POSTMARKET RISK IDENTI-FICATION AND ANALYSIS SYSTEM.—



FDA

Risk Identification and Analysis System:

a systematic and reproducible process to efficiently generate evidence to support the characterization of the potential effects of medical products from across a network of disparate observational healthcare data sources



OMOP to OHDSI





The Observational Health Data Sciences and Informatics (OHDSI) program is a **multistakeholder, interdisciplinary collaborative** to create **open-source** solutions that bring out the value of observational health data through large-scale analytics

OHDSI has established an **international network of researchers and observational health databases** with a central coordinating centre housed at Columbia University





Not pharma funded



International



To improve health

by empowering a community

to collaboratively generate the evidence that promotes

better health decisions and better care





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.



2.

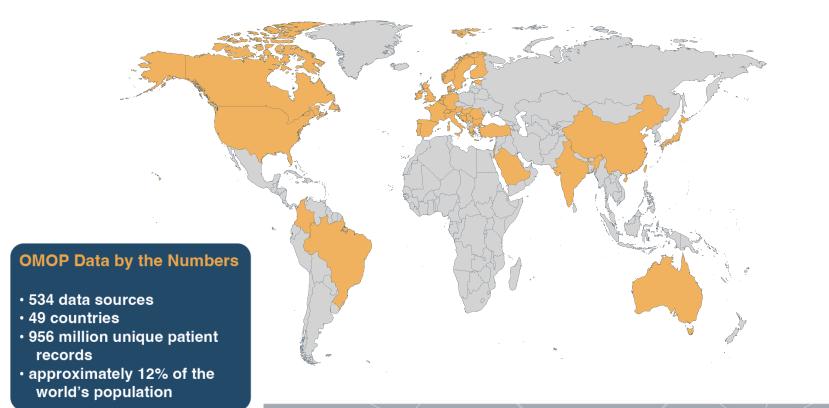
Collaborators

OHDSI By The Numbers

- 3,758 collaborators
- 83 countries
- 21 time zones
- 6 continents
- 1 community

Data Partners





The Author Network



Park T Jin, H Chno-Machado, L OKim, E COO. D etat J OGe, JeKim, W @Song J . Chung, Y Asimwe A Oscillar M. Kim, H CAR I OCha, J OLIN V Viernes, 8 Can Y In an a eTsoucheika, A eRosen Etauner, K •Yoon, J Walbech, J BKry R Thun, S Governa R Shink R 10000 eYoo I eJoo Y BAbn I Belenkava R Oomenech-Femänckez, J Nadal-Almela, S Hwang He Jeon Y Alexa 1 Tak 1 Beuscart J ●Gómez Ádrian, J ●Garcia Garcia E ●Oliver-Garcia, E ·Lamer, A eBarletta V Cee: E Meshoff Del Co. G Sabort-Torses J
 Caparide Redondo, M Californi, A. Crouit 1 Mazzagla, G riedman, C OOT G Massan Strass Carceller, H Montall-Serrano, J Partot A OLapi, P eMa C Dark D Harle Park J Cas D Coloma F Oe La Iglesia-Vaya, M Shee Pasqua Francescont P DuMouch Cohana uchard, M Curve St 00h. S ODuanta D Rassen, J. 61/m eCelertari M Def ako hat K Mathenry N Spotnitz, M #Sun, H Cho, J. Viagoo, T · Drovin Gapne, J
 Perichon, M
 Benichou, J Y-st Averit A Smar, R Catch K CAragón M Ca Klazinga N Banda, J Permandez-Bertolin, S Kent, S Krátka R Shahn Z ·Planting A Cruber, S Oroz-Penoteau, C Tinto C ·Kors J Herps Barela C Molinero, Pressoner, T Boweld, D Bower, D Bower, D Bower, D Bower, D Oliveita. Woore, N etrardin J etraster, LeAtectant, H. etrait H. Bedemann, ettong, NeTaup, S OZhann I A HOOD, Network Court of Olin P Polubriaginof, F Ryan, P Johnson, K GIPTON S Burn, E Carter, W Thurin, N eLi L elghal U eZhan, S Prata-Linbe, A OLU. 8 And Contract N encourses M exclosures S encourses of the second s BOLLET, A OLAL H eMoto Relia OYu. Y OLU, YeChen, J ·Gabetta, M eLarsen, R Oskotsky, B Sec. 1 erini N Dudey. J The Contract Well young state 1 Milety E Contract Well young state 1 Milety E Contract Well young state 1 Milety E Contract A Milety State Jing •Zona N Zhou, X Giangreco, N Binervey, M Wang, Y Volis E latzenia, A Singh, J et orberbaum, T. Tatonatti, N Alloni, A. Espinora J Offen Acalahan, T axe y Bitte A evenner, LLX OVZDAYS D OKhold B Class. C Barat H #Batz. Mol K o Smith D. Medigan, Cl Ba eRuddy, K · Victoritage J eGlicksberg, B eBerlin J eZheng Y Man, K Casirachi, E Ruchspains, V BEatler A eNar v Pasquale, M attrahar Y Description in Mungalt C Stang, P with H ewong. I mweng. Clear N Bate, A Thered approve R #Suchs, B #Worlde, Y Boland M Walentine Gepoeletti Class 2 Rosenberg, M Deer. F eElovici, Y CAMPA W #Shana N Sauer, B Levine M #Zonus M BKint. Cobon K eFrazier, R. Rouhizadeh, M Carrol Tong Denny J est #Orettan D ethi M Conte. Cenot B Variacia ever then. Antinenn •Zampino Dooter 7hat Lingre, T Paktuk M - · #'0416 Tiwati, P P ede Lusignan, S elson, S Cen T OPtat I Wanion, F #Sen, A Chang, X Arror, B •Kulo Corres Cacheco eHarman R. Chairabarti, S Duan, R Rearrissen, L p. Livenage, H Haurster Dabeles, D Cubinet V Fredury. AL-Shuki, S Morris, N .Jang. C Chen W Rusanov, A. eZhang. S Phoker, C Church (OByford, R Ouresh, N •Tandon Watch, F Gold S .ensen, E Hocket, C eHnuby G +ieider. Anand A Carza N Otiobte F OPrudhommeaux, E WWEInns J · hormon Meystre, S Man V Mohan, V Amutha A Waters H. Blaghal, A Hanauer, D evicite R Covinence. 6 Y 8 Coltrin H Zoch M Maver-Devis, E Ganther, A Cimina J Campion, T Cather, M C/Agostino, R Bluilach P Beinecke eWang, P Annalone A Ancies Kansner 1. elGeter P eFischer, P Gathet, F Majeed R Gal W Palet L Claisure, J Gruhi, M Pena Y Pattok
 Shole L Croanback M OX4 Z dillore. ORinner, C event N Adekkanattu, P Paris, N OKaira D Glarzi M Schwachholer, T All antro of Nassirian, A Counters W . Jacobs, J eHaberson A eHenke E Oniel C Vote S Contract H ODeans K OLenert, L Hasse C Bippt P Gruendner, J Croner, R Schottler, J Garstan Benarding J Weatherston, D Guiden, C Gardenheuer, K. Todderroth, D Sedmay, N Conney 1 Maier, C Haverkamp C Tark 1 Boeker, M Storf, H ALIMPIAN G ellatovskiy, A

OHDSI Publications



OHDSI PUBLICATIONS

1. Starg PE, Ryan PB, Raccostn JA, Overhage JM, Hartsema AG, Reich C, Weldob E, Scanecchia T, Woodcock J, Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Ann Intern Med. 2010;15(9):600-8. doi: 10.7326/00034-819-1539-201011020-40010. Publied PMID: 21011580.

 Madigan D, Ryan P. What can we really learn from observational studies?: the need for empirical assessment of methodology for active drug safety surveillance and comsarative effectiveness research. Epidemiology. 2011;22(5):629-31. doi: 10.1097/EDE.0b013e318228ca1d. PubMed PMID: 21811110.

3. Carnahan PM, Moores KG. Mini-Sentinel's systematic reviews of validated methods for identifying health outcomes using administrative and claims data: methods and lessons learned. Pharmacoepidemiol Drug Saf. 2012;21 Suppl 1:82-9. doi: 10.1002/pds.2321. PubMed PMID: 22262596.

4. Overhage JM, Byan PB, Reich CG, Hanzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform As

5. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Med Care. 2012;56 Suppl(0):S21-9. doi: 10.1097/MLR.0b01803182576807. PubMed PMID: 22092254; PubMed Central PMCID: PMCPMC3803892.

 Hech C, Hyan PB, Stang PE, Hocca M. Evaluation of alternative standardized terminologies for medical conductors within a network of coservational networker catabases. J Biomed Inform. 2012;45(4):889-96. Epub 20120607. doi: 10.1016/j.jbi.2012.05.002. PubMed PMID: 22683994.

 Suchard MA, Simpson SE, Zorych I, Ryan P, Madigan D. Massive parallelization of serial inference algorithms for a complex generalized linear model. ACM Trans Model Comput Simul. 2013;23(1). doi: 10.1145/2414416.2414791. PubMed PMID: 25328363: PubMed Central PMCID: PMCPMC4201181.

11. Zonych I, Madigan D, Ryan P, Bate A. Disproportionality methods for pharmacovigilance in longitudinal observational databases. Stat Methods Med Res. 2013;22(1):39-56. Epub 20110830. doi: 10.1177/0962280211403802. PubMed PMID: 21878461.

12. Zhou X, Munugesan S, Bhullar H, Liu Q, Cai B, Wentworth C, Bate A. An evaluation of the THIN database in the OMOP Common Data Model for active drug safety survei lance. Drug Saf. 2013;38(2):119-94. doi: 10.1007/s40284-012-0009-3. PubMed PMID: 23329543.

 Delator FJ, Ryan PB, Soledad Cepeda M, Applying standardized drug terminologies to observational healthcare databases: a case study on opioid exposure. Health Serv Outcomes Res Methods. 2013;13(1):64-7. Epuis 2012;027. doi: 10.1007/s10742-012-1010-1. PubMed FMID: 23398600; PubMed Central BMID: 23398600; PubMed Central BMID: 23398600; PubMed Central BMID: 2339870; PubMed Serveral BMID: 2339870; PubMed Serveral BMID: 2339870; PubMed Serveral BMID: 2339870; PubMed Central BMID: 2339870; PubMed Central BMID: 2339870; PubMed Serveral BMID: 233970; PubMed Serveral BMID: 23397

PMD: 24235108. 15. Madigan D, Ryan PB, Schuemie M. Does design matter? Systematic evaluation of the impact of analytical choices on effect estimates in observational studies. Ther Adv

Drug Saf. 2013;4(2):53-62. doi: 10.1177/2042098613477445. PubMed PMID: 25083251; PubMed Central PMCID: PMCPMC4110833.

 Li X, Hui S, Ryan P, Rosenman M, Overhage M. Statistical visualization for assessing performance of methods for safety surveillance using electronic databases. Phar coepidemiol Drug Saf. 2013;22(5):503-9. Epub 20130214: doi: 10.1002/pds.3419. PubMed PMID: 23408560.

17. Hospat, Dubloadel W, Leffnals P, Bauer-Mehnes, A, Ryan P, Shah HH. Performance of planmacorylpiance signal detection algorithms for the PCPACIBATION proving system. Citic Research and an environmental strategies and an

19. Ogunyemi Ol, Meeker D, Kim HE, Ashish N, Farzaneh S, Boxwala A. Identifying appropriate reference data models for comparative effectiveness research (CER) studies based on data from clinical information systems. Med Care. 2013;51(8 Suppl 3):545-52. doi: 10.1097/MLR.0b013e31829b1e0b. PubMed PMID: 23774519.

 Mardgan D, Ryan PB, Schuernie M, Stang PE, Overhage JM, Hartzema AG, Suchard MA, Duhlouchel W, Berlin JA. Evaluating the impact of atabase heterogeneity or observational study results. Am J Epidemiol. 2013;178(4):845-51. Epub 20130505. doi: 10.1093/ajoi/wt010. PubMed PMID: 2048805; PubMed Central PMCID: PMCP-MC270074.

21. Ryan PB, Madigan D, Stang PE, Schuemie MJ, Hripcsak G. Medication-wide association studies. CPT Pharmacometrics Syst Pharmacol. 2013;2(9):e76. Epub 20130918. doi: 10.1038/psp.2013.52. PubMed PMID: 24448022; PubMed Central PMCID: PMCPMC4026636.

22. Stang PE, Ryan PB, Overhage JM, Schuemie MJ, Hantzema AG, Welebob E. Variation in choice of study design: findings from the Epidemiology Design Decision rv and Evaluation (EDDIE) survey. Drug Saf. 2013;38 Suept 15(15-25; doi: 10.1007/s40284-013-0102-1. PubMed PMID: 24188220.

23. Ryan PB, Schuemie MJ, Welebob E, Duke J, Valentine S, Hartzema AG. Defining a reference set to support methodological research in drug safety. Drug Saf. 2013;36 Suppl 1:S33-47. doi: 10.1007/s40266-013-0097-8. PubMed PMID: 24166222.

24. Hartzema AG, Reich CG, Ryan PB, Stang PE, Madgan D, Welebob E, Overhage JM. Managing data quality for a drug safety surveillance system. Drug Saf. 2013;9/ Suzel 1:549-58. doi: 10.1007/s40264-013-0098-7. PubMed PMID: 24168223.

25. Ryan PB, Schuemie MJ, Gruber S, Zorych I, Madigan D. Empirical performance of a new user cohort method: lessons for developing a risk identification and analysis system. Drug Saf. 2013;36 Suppl 1:559-72. doi: 10.1007/is40284-013-0099-6. PubMed PMID: 24166224.

 Madigan D, Schuemie MJ, Ryan PB. Empirical performance of the case-control method: lessons for developing a risk identification and analysis system. Drug Saf. 2013;86 Suppl 1:S73-82. doi: 10.1007/s40264-013-0105-z. PubMed PMID: 24168225.

27. Suchard MA, Zorych I, Simpson SE, Schuemie MJ, Ryan PB, Madigan D. Empirical performance of the self-controlled case series design: lessons for developing a risk identification and analysis system. Drug Saf. 2013;36 Suppl 1:583-93. doi: 10.1007/s40264-013-0100-4. PubMed PMID: 24166226.

≤ 2013	2014	2015	2016	2017	2018	2019	2020	2021	Thru Sept '22
33	14	21	20	29	36	53	83	103	83
#JoinThe	Journey			e	57				OHDSI.org

22 pages highlighting the 475 publications from our community

OHDSI PUBLICATIONS

494. Microso A, Defado F, Empired assessment of adversitive methods in informing associatily in observation in hearboard stat, BMC Med Nei Methods.
 202222(1):162. De2020703. doi: 10.1146/31247642010639. DAMAdef MIDL 33725111-1 PANAde General MIDL DPADCPRO200712.
 405. Tai MV, Yayi EC, Hun JK, Kin SL, Kin SL, Kin SL, Thearesh ME, Son M, Santon M, Santon

456. Vorisek CN, Lehne M, Klopfenstein SAI, Mayer PJ, Bartschie A, Haese T, Than B, Fast Healthcare Interopenability Resources (FHIR) for Interopenability in Health Research: Systematic Review. JMIR Med Inform. 2022;10(7):x83724. Epub 20220718. doi: 10.2196/05724. PubMed PMID: 35852942; PubMed Central PMIDI: PMICP-M03346569.

457. Kms, Bang, JI, Boo D, Kim B, Choi IY, Ko S, Yoo IR, Kim K, Kim J, Joo Y, Hyoo HG, Paeray JC, Park JM, Jang W, Km B, Ching JN, Yang D, Yoo S, Lae HY. Second primary malignancy risk in hytroid cancer and matched patients with and without radiodeline herapy analysis from the observational health data sciences and informatics. Eur J Naci Med Mol Imaging, 2022;40(1):(3):547–56. Epib 2022041. doi: 10.1007/s000509.02.05779. PubMed PMID: 3082778.

458. Bardenheuer K, Van Speytmosk M, Hague C, Nikal E, Price M. Haematology Outcomes Nethenok in Europe (HONEUR)-A collaborative, interdisophilary platform to hamese her potential of real-world data in hematology. Eur J Haematol. 2022;109(2):134–45. Epub 20220514. doi:10.1111/jb).13780. PubMed PMI0: 3546036. 406.0 Euro E, Ducate Salater J, Fernandez Bedrish B, Reyes C, Konla R, Olemeath A, Rijbback P, Venhamen K, Priesh-Almarka. D. Venous or administ hematologis and death 406.0 Euro E, Ducate Salater J, Ternandez Bedrish B, Reyes C, Konla R, Calmerada A, Rijbback P, Venhamen K, Priesh-Almarka. D. Venous or administ hematologis and death 406.0 Euro E, Ducate Salater J, Fernandez B, Berlosh B, Berlosh P, Venhamen K, Priesh-Almarka. D. Venous or administ hematologis and death 407.0 Euro B, Ducate S, Ducate S, Ducate S, Danie S, Berlosh P, Venhamen K, Priesh-Almarka. D. Venous or administ hematologis and death 408.0 Euro B, Ducate S, Ducate

among COVID-19 cases: a European network orbort study. Lancet Intent Dis. 2022;22(8):1142-52. Epub 20220513. doi: 10.1016/s1473-0096(22)00223-7. PubMed PMID: 35578963; PubMed Central PMCID: PMICPMIC9106320.

460. Lin V, Taouchnika A, Allakhverdiev E, Rosen AW, Gögenur M, Clausen JSR, Birluner KB, Walbech JS, Rijnbeek P, Dnakos I, Gögenur I. Training prediction models for insiVulual inik assessment of postoperative complications after surgery for colorectal cancer. Tech Coloproch. 2022;25(6):565-75. Epub 20220520. doi: 10.1007/s10151-022-002824-x. PubMed MIDI: 3555971.

461. Bistuner KB, Rosen AW, Tsouchnika A, Wabech JS, Gógenur M, Lin VA, Clausen JSR, Gógenur L. Developing prediction models for short-term mortality after surgery for colonical canner using a Damien national quality assurance database. Int J Colorectal Dis. 2002;27(8):1883–43. Epub 20220718. doi: 10.1007/s00384-022-040274. PubMed PMID: 3984/995.

462. Lamer A, Moussa MD, Marolly R, Logier R, Vallet B, Tavernier B. Development and usage of an anesthesia data warehouse: lessons learnt from a 10-year project. J Clin Monit Comput. 2022. Epub 20220806. doi: 10.1007/s10877-022-00896-y. PubMed PMID: 35933465.

464. Glangreco NP, Tatonetti NP. A database of pediatric drug effects to evaluate ontogenic mechanisms from child growth and development. Med (N Y). 2022;3(8):579-95.e7. Epub 20220624. doi: 10.1016/j.medj.2022.06.001. PubMed PMID: 35752163; PubMed Central PMCID: PMCPMC9378670.

465 Holmman A. Ruanga C. Maanubagawa G. Magori S, Senshali M, Kitaban L, Ji Kanan L, Li Manan A, Uhanateshma JM, Katabashta JJ, Olivana G, Alanosh B, Talazen L, Shangi F, Haushan JM, Katabashta JJ, Divensign attifaid in Lipescamp F, Hauphang J, Burkin B, Bahol D, Bulmana G, Janosh B, Talazen C, Nuorkinyesu K, Bunnyi F, Kauoninana D, Katabashta JJ, Linversign attifaid intiligence and alas interne techniques in humorizing, shaning accessing and analysing SARI COVIACOVID 18 data in Reanda (LAIGNAH Popel); study alega and nationala. BM: Med Hom: Denis Mai. 2022;22(1):214. Epub 2020512: doi: 10.1180/s12911-022-01805-0. PubMed DND: S5602305; PubMed Cemai PMCID: PMCIP. IO20272051.

466. Swerdel JN, Schuemie M, Murray G, Ryan PB. PheValuator 2.0: Methodological improvements for the PheValuator approach to semi-automated phenotype algorithm evaluation. J Biomed Inform. 2022;104177. Epub 20220819. doi: 10.1016/j.jbi.2022.104177. PubMed PMID: 35995107.

487 Delawente G, Willame R, Biganoc A, Bytere R, Fortes A, Tanzy RMM, Anerd SN, Bealery D, Karpy R, Asbard, Beestan S, Guora RA, Cavee R, Toose R, Toose R, Toose N, Toose R, Toose R, Toose N, Toose R, Toose R,

468. Williams RD, Raps JM, Rijcheek PR, Ryan PB, Prieto-Ahambra D. 90-Day all-cause mortality can be predicted following a total innee replacement: an international, network which to develop and validate a prediction model. Knee Surg Sports Traumatel Arthrosc. 2022;20(9):3068-75. Epub 2021;208. doi: 10.1007/a00167-621-061799-y. P-Abder 5M0: 32870731.

469. Abeysinghe R, Black A, Kaduk D, Li Y, Reich C, Davydov A, Yao L, Cui L. Towards quality improvement of vaccine concept mappings in the OMOP vocabulary with a semi-automated method. J Biomed Inform. 2022;134:104162. Eoub 20220825. doi: 10.1016/j.bi.2022.104162. PubMed PMID: 36029954.

470. Almeida JR, Barraca JP, Oliveira JL. Preserving Privacy when Querying OMOP CDM Databases. Stud Health Technol Inform. 2022;298:163-4. doi: 10.3233/sht/220930. PubMed PMID: 36073478.

471. Williams PD, Reps JM, Riybeek PR, Ryan PB, Prieto-Ahambra D. 90-Day all-cause mortality can be predicted following a total ince replacement: an international, network study to develop and validate a prediction model. Knee Surg Sports Traumatol Arthrosc. 2022;0(0):3088-75. Epub 20211208. doi: 10.1007/s00167-021-00799-y PubMed PMID: 3047791.

472, Xlao G, Plaff E, Prudhommeaux E, Booth D, Sharma DK, Huo N, Yu Y, Zong N, Ruddy KJ, Chute CG, Jiang G. FHIR-Ontop-OMOP: Building Clinical Knowledge Graphs in FHIR RDF with the OMOP Common Data Model. J Biomed Inform. 2022-104201. Epub 20220908. doi: 10.1016/j.jbi.2022.104201. PubMed PMID: 36089199.

473. Zhang L, Wang Y, Schuemie MJ, Biei DM, Hriposak G. Adjusting for indirectly measured confounding using large-scale propensity score. J Biomed Inform. 2022;104204 Epub 20220912. doi: 10.1016/j.jpi.2022.104204. PubMed PMID: 36108916.

474. Castraro VG, Spohitz M, Waldman GJ, Joiner EF, Choi H, Ostropolets A, Natanajan K, McKham GM, Ottman R, Neugat AJ, Heposak G, Yoongerman BE. Identification of patients with drug resistant egilepsy in electronic medical record data using the Observational Medical Outcomes Partnership Common Data Model. Epilepsia. 2022 Epub 2022014. doi: 10.1111/jeq.11420...PubMed PMID: 36105877.

173. Nondowend M, Price BB, Prostnikid JZ, Burnell HT, Vest MT, Anzahon AJ, Happe J, Kimble WD, Moradi H, Hendricks B, Santangele BL, Hodder SL. An ordinal severit scale for COVID-19 retrospective studies using Electronic Health Record data. JAMA Open. 2022;5(3):ooac086. Epub 20220709. doi: 10.1093/jamiacpeniooac086. PubMe MMD: 59911666. PubMed Cereral PMICIP. PMICPM027199.



OHDSI Publications





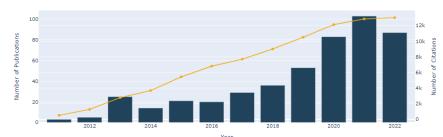
Community Dashboard Dashboards -

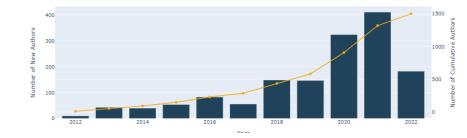
Publication Analysis

PubMed Publication Tracking highlights scholarship generated using the OMOP Common Data Model, OHDSI tools, or the OHDSI network. These publications represent scientific accomplishments across areas of data standards, methodological research, open-source development, and clinical applications. We provide the resource to search and browse the catalogue of OHDSI-related publications by date, author, title, journal, and SNOMED terms. We monitor the impact of our community using summary statistics (number of publications and citations), and the growth and diversity of our community with the number of distinct authors. Searches for new papers are performed daily, and citation counts are updated monthy.

OHDSI Publications & Cumulative Citations

New and Cumulative OHDSI Researchers

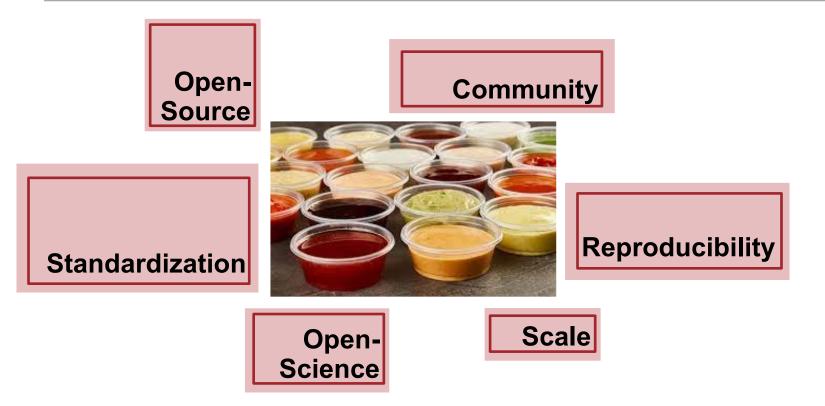




Explore our community progress: http://dash.ohdsi.org

The Secret Sources





Summary and Segue



- OHDSI largest and fastest growing community for RWE
- Because of
 - Standardization
 - Open Source
 - Open Science
 - Reproducibility
 - Community
 - Scale
- You should join, too

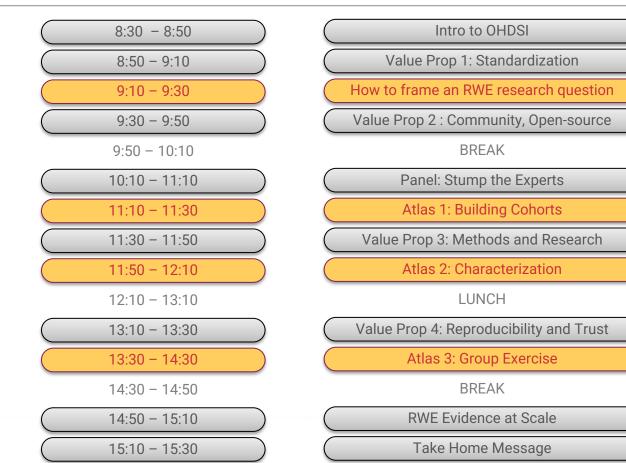


Join the Journey!



Agenda





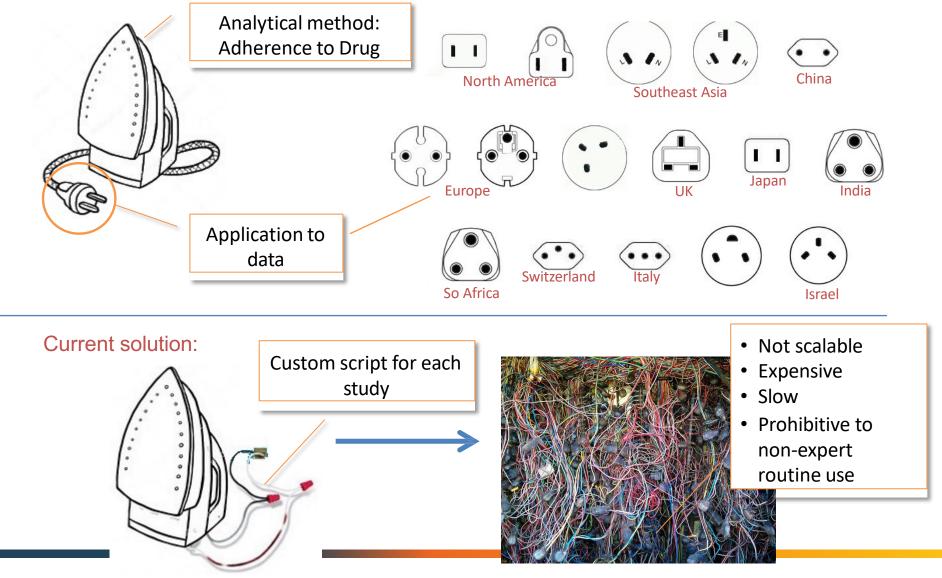


OHDSI Standardization of Evidence Generation



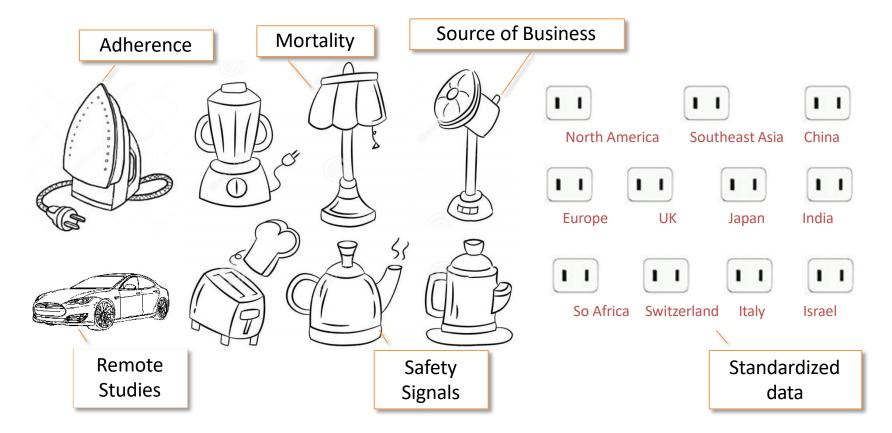
Current Approach: "One Study – One Script"

"What's the adherence to my drug in the data assets I own?"



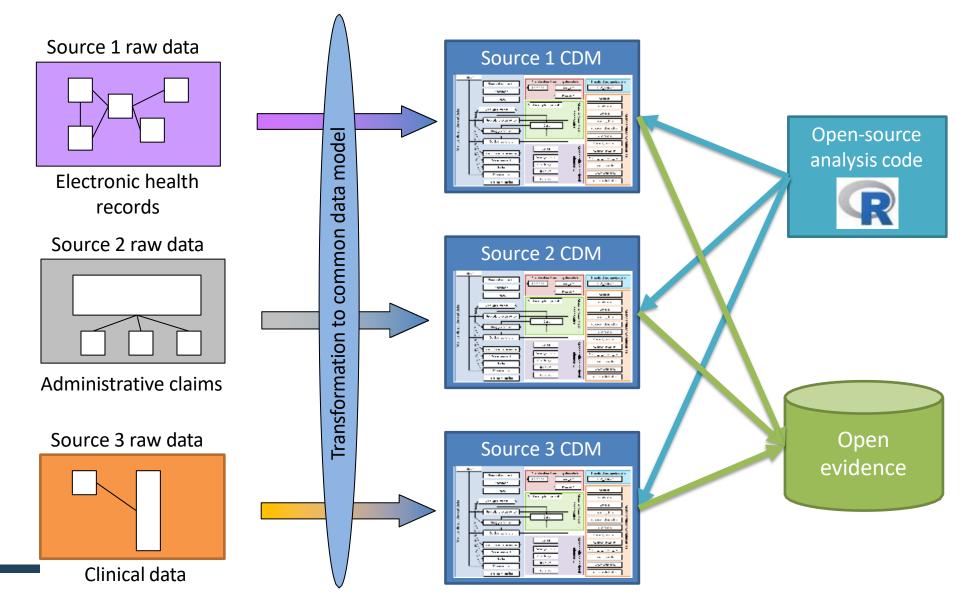


Solution: Standardized Data and Analytics



- 1. ATLAS, Remote Studies
 - Standard Cohorts
 - Standardized Analytics
- 2. OMOP CDM
 - Standardized Format
 - Standardized Coding

Common data model can enable standardized analytics across a distributed data network





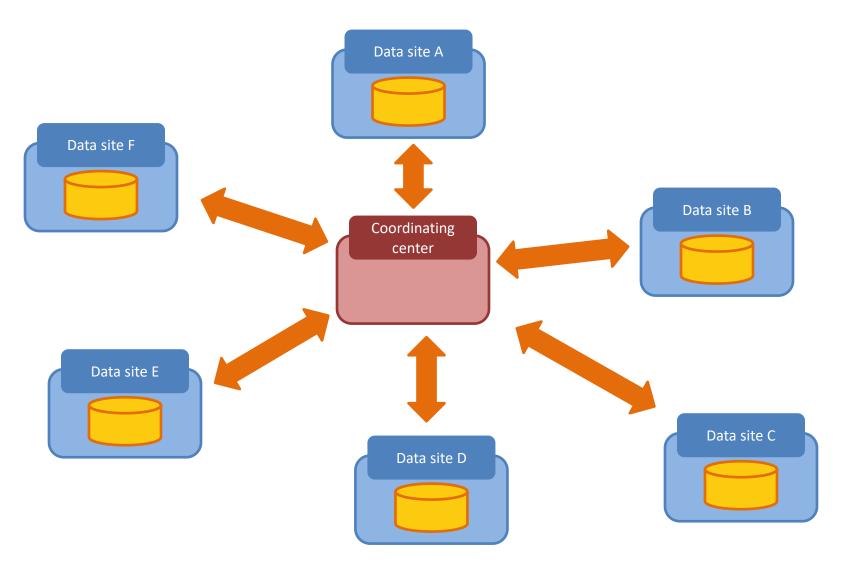
Research Across Distributed Research Networks

Traditional way:

- 1. Share data
- 2. Harmonize data
- 3. Then analyze

OHDSI way:

- 1. Leave data where it is
- 2. Harmonize each site's data to OMOP CDM
- 3. Share aggregated statistics for analysis





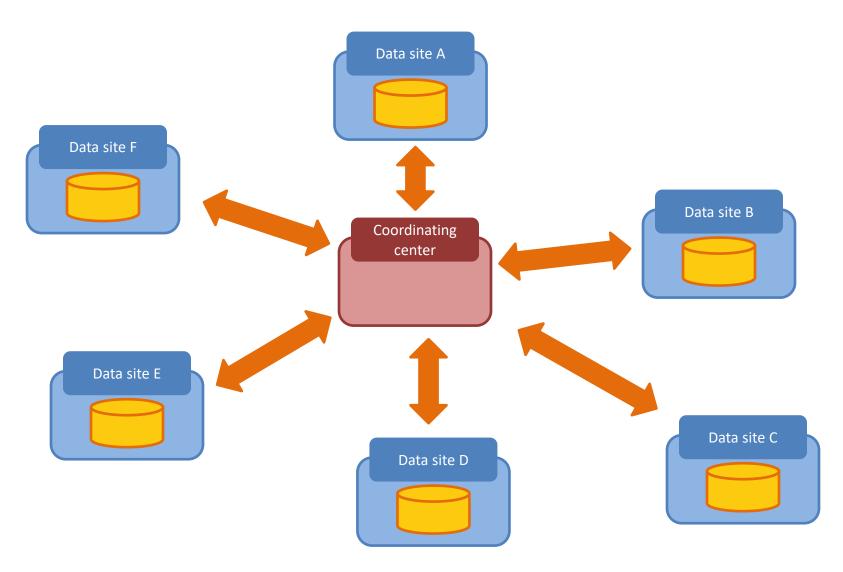
Research Across Distributed Research Networks

Same across sites:

- Common Data Model
- Standardized Vocabularies
- Phenotypes/Cohorts
- Analysis/Methods
- Evidence generation

Different across sites:

- Health care system
- Data capture process
- Source coding systems
- ETL
- Database platform





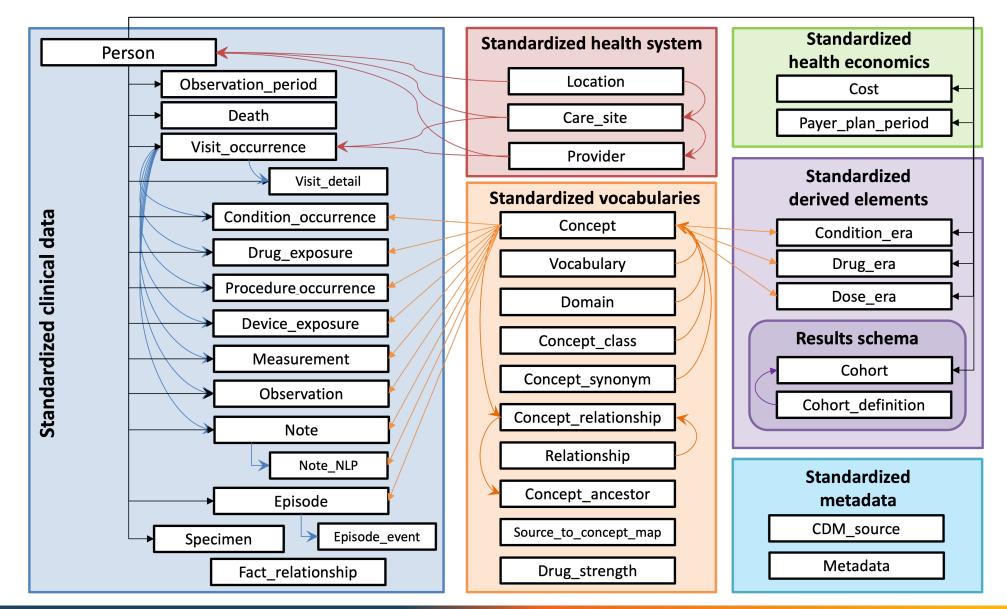
OMOP Common Data Model

- Components
 - Schema tables where you put data
 - Vocabulary what codes go in the table
 - Conventions how to store data

- Open committee structure to govern it
 - Contracted vocabulary maintenance



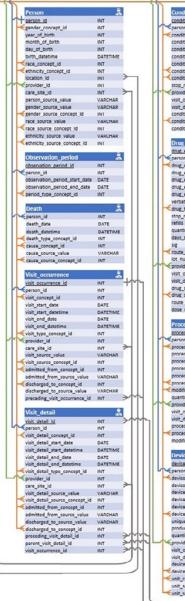
OMOP Common Data Model

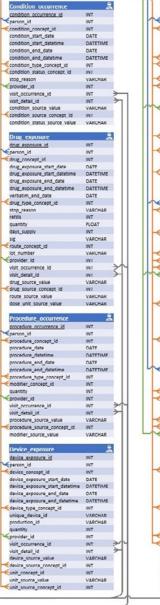


OMOP Common Data Model v5.4



roorder, Id INT roorder, Id INT VARCHAR pl VARCHAR pl VARCHAR peclaty, concept id INT mear of birth INT mear of birth INT mear of birth INT mear of birth INT memory source value VARCHAR peclaty, source concept id INT concerner, it VARCHAR concerner, it VARCHAR the of service concept id INT concerner, it VARCHAR the of service vAR	+
bi VARCHAR Bi VARCHAR Be VARCHAR Mediatry Concept, Id INT	+
is VARCHAR Socially concept Id INT Fair of Dirth INT INT Sociality Jource value VARCHAR Sociality Jource value Sociality Jource value VARCHAR Sociality Jource value VARCHAR Sociality INT Fair Site Id INT Sociality VARCHAR Sociality	*
peciality concept id init are size id init ears of birth init ears of birth init ears of birth init initial concept id init ender, source value VARCHAR enders, source concept id init enders source concept id init ears size, name VARCHAR ears size, name VARCHAR issee of service concept id init ears size, name VARCHAR issee of service concept id init ears size, name VARCHAR issee of service concept id init ears size, of service concept id init ears size of service concept id init ears size, of service concept id init ears	*
are size Id INT see of birth INT wrider_concept_Id INT wrider_concept_Valve VARCHAR peckilly_source_valve VARCHAR peckilly_source_valve VARCHAR medies_source_valve VARCHAR medies_source_valve VARCHAR medies_source_valve VARCHAR medies_source_valve VARCHAR medies_of_source_valve VARCHAR mediase_of_source_valve VARCHAR	
Rei of John INI Rei of John INI rovide Jource, value VARCHAR pockalis, yource, value VARCHAR pockalis, yource, value VARCHAR moder Jource value VARCHAR ender Jource value VARCHAR are Jite Init are jite Init are jite Jource value VARCHAR init are jite Jource value VARCHAR varchar conton of Service Jource value VARCHAR po VARCHAR po VAR	
involde_publick_source_value VARCHAR peciality_source_concept_id INY are_site	
varchar pociality jource concept id initian ender source value varchar ender source concept id initian ender source value varchar ender source value varchar p varchar p varchar	
opeciality source, concept Jd INI meder, source, value VARCHAR ender, source, concept Jd INI are, site, Id INI are, site, or concept Jd INI watchar, are of service, concept, Jd INI are, site, or concept, Jd INI watchar, J watchar, or concept, Jd INI watchar, or concept, Jd INI watchar, or concept, Jd INI watchar, J watchar, J watchar, or concept, Jd INI watchar, J watchar, J	
energe source value vaccuax ender source concept id int care_site_id int care_site_id int vaccuax enders_source_value vaccuax idexe_site_id int care site_id int care site_id int care site_id int vaccuax idexe_site_source_value vaccuax idexes_i	
ender source concept la INV are_she is are_she is are_she is are she hame VARCHAR INV Are she hame INV INV Are she hame INV INV INV Are she hame INV INV Are she INV INV INV Are she INV INV INV Are she INV INV Are she INV INV INV INV Are she INV INV INV INV INV INV INV INV	
are site in any set of a set of any set of a	
rer site, lot init inter site, name VAACHAN lace of service, concept jol init acce of service, concept jol init init init init init accel of service, value VAACHAN varchan v	
var site name varchak ine of service concept id init castion_id init acation_id init init castion_id init init castion_id init varchak)
vise of service concept id init conton_id init are_set_source_value VARCHAR vare_set_source_value VARCHAR conton id init ocation id init oddrest_1 VARCHAR of varCHAR varCHAR p VARCHAR varCHA)
cación_id int acación_id VARCHAR lace_of_service_source_value VARCHAR costion VARCHAR costion id INT ddress_1 VARCHAR ddress_2 VARCHAR tate VARCHAR p VARCHAR p VARCHAR p VARCHAR contro_source_value VARCHAR toutro_concept_d INT vARCHAR	3
Are Lefe Jource, value VARCHAR isce, of service, source, value Cention C	3
Islace of service source value VARCHAR OCation Value VARCHAR VARCHAR VARCHAR VARCHAR VARCHAR VARCHAR vare VARCHAR vare VARCHAR vare VARCHAR vare v	3
Intrinoid Int ddrest_1 VBCHAB ddrest_2 VARCHAB thy VARCHAB p VARCHAB p VARCHAB p VARCHAB coator_pourte_value VARCHAB coator_pourte_value VARCHAB studie FLOAT	\$
xxx10x.id IAT xx20x.id IAT ddrest_1 VAECHAB ddrest_2 VAECHAB ty VAECHAB p VAECHAB p VAECHAB xxmm VAECHAB xxatom_uourte_value VAECHAB xxatom_uourte_value VAECHAB xxatom_uouth_concept_id IAT xxatom_uouth_concept_id IAT textue FLOAT	-
ddevsr_1 VARCHAR ddrest_2 VARCHAR try VARCHAR trate VARCHAR trate VARCHAR trate VARCHAR trate VARCHAR trate VARCHAR contry_concept_id INT contry_concept_id INT contry_concept_id INT contry_concept_id INT	
ddresi_2 VARCHAR tarse VARCHAR tarse VARCHAR ip VARCHAR contry VARCHAR contry Concept_d INT contry_concept_d INT contry_concept_d VARCHAR totude FLOAT	
thy VARCHAR tane VARCHAR top VARCHAR ownry VARCHAR colon, Source_value VARCHAR ountry_concept_id INT country_concept_id INT totude FLOAT	
ip VARCHAR county VARCHAR counto_source_value VARCHAR counto_concept_id INT ounto_concept_id INT VARCHAR attude FLOAT	
vointy VARCHAR ocation_source_value VARCHAR ountry_concept_id INT iountry_source_value VARCHAR attude FLOAT	
ocation_source_value VARCHAR country_concept_id INT country_source_value VARCHAR attude FLOAT	
country_concept_id INT country_source_value VARCHAR atitude FLOAT	
country_source_value VARCHAR atitude FLOAT	
atitude FLOAT	
Metadata	i
netadata id INT	
netadata_concept_id INT	
netadata_type_concept_id INT	
ame VARCHAR	
alue_as_string VARCHAR alue_as_concept_id INT	
alue_as_concept_id INT alue_as_number FLOAT	
netadata_date DATE	
netadata_datetime DATETIME	21
dm_source	1
dm_source_name VARCHAR dm_source_abbreviation VARCHAR	
dm_holder VARCHAR	
ource_description VARCHAR	
ource_documentation_reference VARCHAR	Í.
dm_etl_reference VARCHAR	
ource_release_date DATE	
dm_release_date DATE	i.
dm_version VARCHAR	
dm_version_concept_id INT	





Measurement	2	
easurement id	INT	
erson_id	INT	
easurement_concept_id	INT	
neasurement_date neasurement_datetime	DATE	
	DATETIME	
measurement_time	VARCHAR	
measurement_type_concept_id	INT	
operator_concept_id	INT	
value_as_number	FLOAT	
value_as_concept_ld	INT	
unit_concept_id	INT	
ange_low	FLOAT	
range_high	FLOAT	
provider_id	INT	~
isit_occurrence_id	INT	~
isit_detail_id	INT VARCHAR	-
measurement_source_value measurement_source_concept_id	INT	
unit_source_value	VARCHAR	
	INT	
unit_source_concept_id value_source_value	VARCHAR	
measurement_event_ld	INT	
measurement_event_id meas_event_field_concept_id	INT	
meas event new concept 10	-	
Observation		
	INT	
observation id	INT	
person_id observation_concept_id	INT	
	DATE	
observation_date observation_datetime	DATE	
	INT	
observation_type_concept_id value_as_number	FLOAT	
value_as_number	VARCHAR	
value_as_string value_as_concept_id	INT	
qualifier_concept_id	INT	
unit_concept_id	INT	
provider_id	INT	
visit_occurrence_id	INT	2
visit_detail_id	INT	5
observation_source_value	VARCHAR	1
observation_source_concept_id	INT	
unit_source_value	VARCHAR	
qualifier_source_value	VARCHAR	
value source value	VARCHAR	
observation_event_ld	INT	
bs_event_field_concept_id	INT	
lote	2	
ate id	INT	>
erson_id	INT	
note_date	DATE	
note_datetime	DATETIME	
note_type_concept_id	INT	
ote_class_concept_id	INT	
ote title	VARCHAR	
note_text	VARCHAR	
encoding_concept_id	INT	
anguage_concept_id	INT	
rovider_id	INT	
isit_occurrence_id	INT	*
visit_detail_id	INT	+
note_source_value	VARCMAR	
note_event_ld	INT	
note_event_field_concept_id	INT	
		-
Note_nip		5
note nip id	INT	
	INT	*
note_id		
note_id section_concept_id	INT	
note_id	VARCHAR	
note_id section_concept_id snippet	VARCHAR	
note_id section_concept_id snippet offset lexical_variant	VARCHAR VARCHAR VARCHAR	
note_id section_concept_id snippet offset axical_variant note_nip_concept_id	VARCHAR	
note_id section_concept_id offset lexical_variant note_nip_concept_id note_nip_cource_concept_id	VARCHAR VARCHAR VARCHAR INT	
note_id section_concept_id offset lexical_variant note_nip_concept_id note_nip_cource_concept_id	VARCHAR VARCHAR VARCHAR INT	
note_id section_concept_id snippet offset	VARCHAR VARCHAR VARCHAR INT	
note_id section_concept_id sinpet difuet lexical_variant note_nip_concept_id note_nip_source_roncept_id nip_system	VARCHAR VARCHAR VARCHAR INT VARCHAR	
note_id section_concept_id offsat lexical_variant note_nip_concept_id note_nip_source_concept_id nip_system nip_system	VARCHAR VARCHAR VARCHAR INT INT VARCHAR DATE	
note_id section_concept_id offset sector_torept_id note_nip_concept_id nip_system nip_system nip_date mip_datetime	VARCHAR VARCHAR INT INT VARCHAR DATE DATETIME	

pecimen	
ecimen ld	INT
rson_id	INT
cimen_concept_id	INT
cimen_type_concept_id cimen_date	DATE
scimen_datetime	DATETIME
antity	FLOAT
t concept id	INT
itomic_site_concept_id	INT
ease_status_concept_ld	INT
cimen_source_ld	VARCHAR
cimen_source_value	VARCHAR
t_source_value	VARCHAR
ease status source value	VARCHAR
ct_relationship	4
ct_relationship nain_concept_id_1	INT
t_id_1	INT
main_concept_id_2	INT
t_id_2	INT
tionship_concept_id	INT
ડા	
<u>t id</u> t_event_id	INT
t_event_id it_domain_id	VARCHAR
t_type_concept_id	INT
rency_concept_id	INT
al_charge	FLOAT
al_cost	FLOAT
al_paid	FLOAT
d_by_payer	FLOAT
d_by_patient	FLOAT
d_patient_copay	FLOAT
d_patient_coinsurance d_patient_deductible	FLOAT
id_by_primary	FLOAT
id_ingredient_cost	FLOAT
d dispensing fee	FLOAT
ver_plan_period_id	INT
ount_allowed	FLOAT
enue_code_concept_id	INT
enue_code_source_value	VARCHAR
concept_id t_source_value	INT VARCHAR
yer_plan_period	INT
ver plan period id rson_id	INT
ver_plan_period_start_date	DATE
er_plan_period_end_date	DATE
ver_concept_id	INT
er_source_value	VARCHAR
ver_source_concept_id	INT
n_concept_id	INT
n_source_value	VARCHAR
n_source_concept_id	INT
nsor_concept_id	INT
ensor_source_value	VARCHAR
and course concept 14	VARCHAR
insor_source_concept_id	VARUMAR
nisor_source_concept_id nily_source_value	INT
nsor_source_concept_id nily_source_value p_reason_concept_id p_reason_source_value	INT VARCHAR

Condition_era	1
condition era id	INT
<pre>person_id</pre>	INT
<pre>condition_concept_id</pre>	INT
condition_era_start_date	DATE
condition_era_end_date	DATE
condition_occurrence_count	INT
-	
Drug_era	5,7
drug era id	INT
erson_id	INT
<pre> drug_concept_id </pre>	INT
drug_era_start_date drug_era_end_date	DATE
drug_era_end_date drug_exposure_count	INT
gap_days	INT
0-7	
Dose era	
dose era id	INT
<pre>person_id</pre>	INT
drug_concept_id	INT
unit_concept_id	INT
dose_value	FLOAT
dose_era_start_date	DATE
dose_era_end_date	DATE
Episode	1
episode id	INT
episode id person id	INT
episode_concept_id	INT
episode_start_date	DATE
episode_start_datetime	DATETIME
episode_start_oateome episode_end_date	DATE
episode_end_datetime	DATE
episode_eno_datetime episode_parent_id	INT
episode_number	INT
episode_object_concept_id	INT
episode_type_concept_id	INT
episode_source_value	VARCHAR
<pre>episode_source_concept_id</pre>	INT
Episode_event	14
episode_id	INT
event_id	INT
episode_event_field_concept_id	INT
Cohort	0
cohort_definition_id	INT
subject_id	INT
cohort start date	DATE
cohort_end_date	DATE
Cohort_definition	14
cohort_definition_id	INT
cohort_definition_name	VARCHAR
cohort_definition_description	VARCHAR
<pre> definition_type_concept_id</pre>	INT
cohort_definition_syntax	VARCHAR
	INT
<pre>subject_concept_id cohort_initiation_date</pre>	DATE

Legend	
Clinical data tables	2
Health system data tables	Ś
Health economics data tables	
Standardized derived elements	- 0
Metadata tables	i
Vocabulary tables	m
Primary key	



OHDSI's Standardized Vocabularies

		1		
CONCEPT_ID	313217	←	Primary key	
CONCEPT_NAME	Atrial fibrillation	<	English description	1
DOMAIN_ID	Condition	<	Domain	
VOCABULARY_ID	SNOMED	<	Vocabulary	
CONCEPT_CLASS_ID	Clinical Finding	<	Class in vocabulary	
STANDARD_CONCEPT S		←	Standard, Source	
CONCEPT_CODE	49436004	*	of Classification	
VALID_START_DATE	01-Jan-1970	K	Code in vocabulary	/
VALID_END_DATE	31-Dec-2099	\leftrightarrow	Valid during time	
INVALID_REASON		4	interval	

Standard representation of vocabulary concepts in the OMOP CDM. The example provided is the CONCEPT table record for the SNOMED code for Atrial Fibrillation.



OHDSI's Standardized Vocabularies

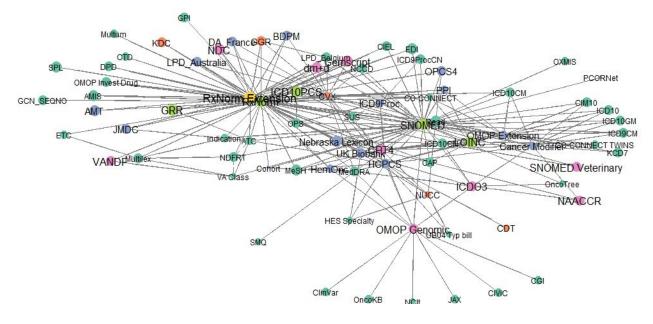
- 142 Vocabularies across 44 domains
 - MU3 standards: SNOMED, RxNorm, LOINC
 - Disparate sources: ICD9CM, ICD10(CM), Read, NDC, Gemscript, CPT4, HCPCS...
- >11 million concepts
 - 3.6 million standard concepts5.1 million source codes847k classification concepts
- 82 million concept relationships
- 88 million ancestral relationships

Often referred to as "The Vocabulary"

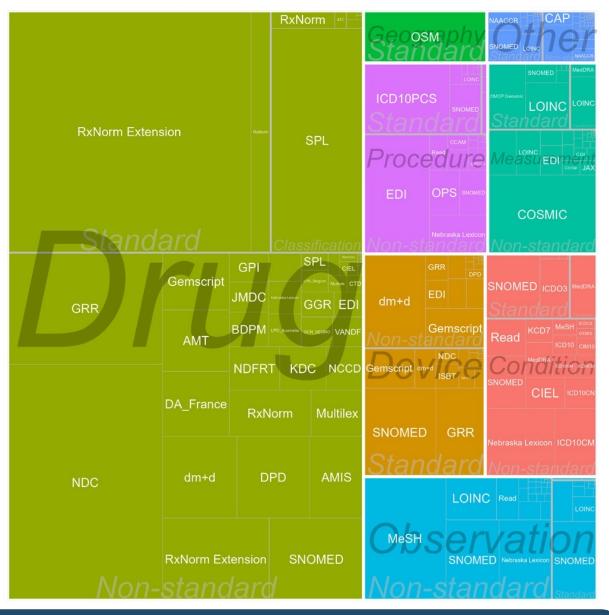
Publicly available at: https://athena.ohdsi.org/



OHDSI Standardized Vocabularies



This network diagram shows the relationships between vocabularies. Nodes are vocabularies, sized by the number of concepts. Edges show connections between concepts within vocabularies.



This treemap shows all concepts in the OHDSI vocabularies, organized by domain (color) and vocabularies (boxes sized by the number of concepts).



Standardized Analysis

OHDSI standardized analytics use cases:

- Characterization
- Population-level estimation
- Patient-level prediction



HADES (formally known as the OHDSI Methods Library) is a set of open-source R packages for large scale analytics, designed specifically for direct interaction with the OMOP CDM.



Standardized Analysis

HADES

HADES is a set of open source R packages for large scale analytics, including population characterization, population-level causal effect estimation, and patient- level prediction.

The packages offer R functions that together can be used to perform an observational study through the full journey from data to evidence, including data manipulation, statistical modeling, and results generation with supporting statistics, tables and figures.

Each package includes functions for specifying and subsequently executing multiple analyses efficiently. HADES supports best practices for use of observational data as learned from previous and ongoing research, such as transparency, reproducibility, as well as measuring of the operating characteristics of methods in a particular context and subsequent empirical calibration of estimates produced by the methods.

Population-Level Estimation

CohortMethod

CohortMethod is an R package for performing new-user cohort studies in an observational database in the OMOP Common Data Model.

EvidenceSynthesis

This R package contains routines for combining causal effect estimates and study diagnostics across multiple data sites in a distributed study. This includes functions for performing meta-analysis and forest plots.

SelfControlledCaseSeries

SelfControlledCaseSeries is an R package for performing Self-Controlled Case Series (SCCS) analyses in an observational database in the OMOP Common Data Model.

SelfControlledCohort

This package provides a method to estimate risk by comparing time exposed with time unexposed among the exposed cohort.

Patient-Level Prediction/Characterization

PatientLevelPrediction

PatientLevelPrediction is an R package for building and validating patient-level predictive models using data in the OMOP Common Data Model format.

DeepPatientLevelPrediction

DeepPatientLevelPrediction is an R package for building and validating deep learning patient-level predictive models using data in the OMOP Common Data Model format and OHDSI PatientLevelPrediction framework.

EnsemblePatientLevelPrediction

EnsemblePatientLevelPrediction is an R package for building and validating ensemble patient-level predictive models using data in the OMOP Common Data Model format. The package expands the OHDSI R PatientLevelPrediction package to enable ensemble learning.

Characterization

Characterization is an R package for performing characterization of a target and a comparator cohort.



Standardized Analysis

Cohort Construction

CAPR

The goal of Capr, pronounced 'kay-pr' like the edible flower, is to provide a language for expressing OHDSI Cohort definitions in R code. OHDSI defines a cohort as "a set of persons who satisfy one or more inclusion criteria for a duration of time" and provides a standardized approach for defining them (Circe-be). Capr exposes the standardized approach to cohort building through a programmatic interface in R which is particularly helpful when creating a large number of similar cohorts. Capr version 2 introduces a new user interface designed for readability with the goal that Capr code being a human readable description of a cohort while also being executable on an OMOP Common Data Model.

CirceR

A R-wrapper for Circe, a library for creating queries for the OMOP Common Data Model. These gueries are used in cohort definitions (CohortExpression) as well as custom features (CriteriaFeature). This package provides convenient wrappers for Circe functions, and includes the necessary Java dependencies.

CohortDiagnostics

CohortDiagnostics is an R utility package for the development and evaluation of phenotype algorithms for OMOP CDM compliant data sets. This package provides a standard, end to end, set of analytics for understanding patient capture including data generation and result exploration through an R Shiny interface. Analytics computed include cohort characteristics, record counts, index event misclassification, captured observation windows and basic incidence proportions for age, gender and calendar year. Through the identification of errors, CohortDiagnostics enables the comparison of multiple candidate cohort definitions across one or more data sources, facilitating reproducible research.

CohortExplorer

This software tool is designed to extract data from a randomized subset of individuals within a cohort and make

it available for exploration in a 'Shiny' application environment. It retrieves date-stamped, event-level records from one or more data sources that represent patient data in the Observational Medical Outcomes Partnership (OMOP) data model format. This tool features a user-friendly interface that enables users to efficiently explore the extracted profiles, thereby facilitating applications, such as reviewing structured profiles. The output of this R-package is a self-contained R shiny that contains person-level data for review.

CohortGenerator

This R package contains functions for generating cohorts using data in the CDM.

PheValuator

The goal of PheValuator is to produce a large cohort of subjects each with a predicted probability for a specified health outcome of interest (HOI). This is achieved by developing a diagnostic predictive model for the HOI using the PatientLevelPrediction (PLP) R package and applying the model to a large, randomly selected population. These subjects can be used to test one or more phenotype algorithms.

PhenotypeLibrary

The OHDSI community has developed a publicly accessible, version-controlled Phenotype Library to guide real-world evidence towards the FAIR principles: Findability, Accessibility, Reproducibility, and Interoperability. This library aims to foster the submission and retrieval of high-quality cohort definitions, cataloging of metadata, attribution and promotion of discovery and reuse in scientific research. Within the OHDSI Phenotype Library (OHDSI PL), each entry represents a unique cohort definition identifiable by a stable, externally referenceable ID. Comprehensive metadata about each cohort definition is cataloged and made searchable for researchers.Content in the library is subject to version control, with each version is assigned a specific DOI.

Evidence Quality

Achilles

Automated Characterization of Health Information at Large- DataQualityDashboard (DQD) is an R package for Scale Longitudinal Evidence Systems (ACHILLES) Achilles provides descriptive statistics on an OMOP CDM database. ACHILLES currently supports CDM version 5.3 and 5.4.

Data Quality Dashboard

exposing and evaluating observational data guality. This package runs a series of data guality checks against an OMOP CDM instance. It systematically runs the checks. evaluates each check against a pre-specified threshold. and then communicates what was done in a transparent and easily understandable way.

Evidence Quality

EmpiricalCalibration

This R package contains routines for performing empirical calibration of observational study estimates. By using a set of negative control hypotheses we can estimate the empirical null distribution of a particular observational study setup. This empirical null distribution can be used to compute a calibrated p-value, which reflects the probability of observing an estimated effect size when the null hypothesis is true taking both random and systematic error into account, as described in the paper Interpreting observational studies: why empirical calibration is needed to correct p-values.

Also supported is empirical calibration of confidence intervals, based on the results for a set of negative and positive controls, as described in the paper Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data.

Method Evaluation

This R package contains resources for the evaluation of the performance of methods that aim to estimate the magnitude (relative risk) of the effect of a drug on an outcome. These resources include reference sets for evaluating methods on real data, as well as functions for inserting simulated effects in real data based on negative control drug-outcome pairs. Further included are functions for the computation of the minimum detectable relative risks and functions for computing performance statistics such as predictive accuracy, error and bias.

Supporting Packages

Andromeda

AsynchroNous Disk-based Representation of MassivE

DAta (ANDROMEDA): An R package for storing large data objects. Andromeda allow storing data objects on a local drive, while still making it possible to manipulate the data in an efficient manner.

BiaKNN

An R package implementing a large scale k-nearest neighbor (KNN) classifier using the Lucene search engine.

BrokenAdaptiveRidge

BrokenAdaptiveRidge is an R package for performing L 0based regressions using Cyclops.

Cyclops

Cyclops (Cyclic coordinate descent for logistic, Poisson and survival analysis) is an R package for performing large scale regularized regressions.

DatabaseConnector

This R package provides function for connecting to various DBMSs. Together with the SglRender package, the main goal of DatabaseConnector is to provide a uniform interface across database platforms: the same code should run and produce equivalent results, regardless of the database back end.

Eunomia is a standard dataset in the OMOP (Observational Medical Outcomes Partnership) Common Data Model (CDM) for testing and demonstration purposes. Eunomia is used for many of the exercises in the Book of OHDSI. For functions that require schema name, use 'main'.

FeatureExtraction

An R package for generating features (covariates) for a cohort using data in the Common Data Model.

Hydra

Eunomia

An R package and Java library for hydrating package skeletons into executable R study packages based on specifications in JSON format.

IterativeHardThresholding

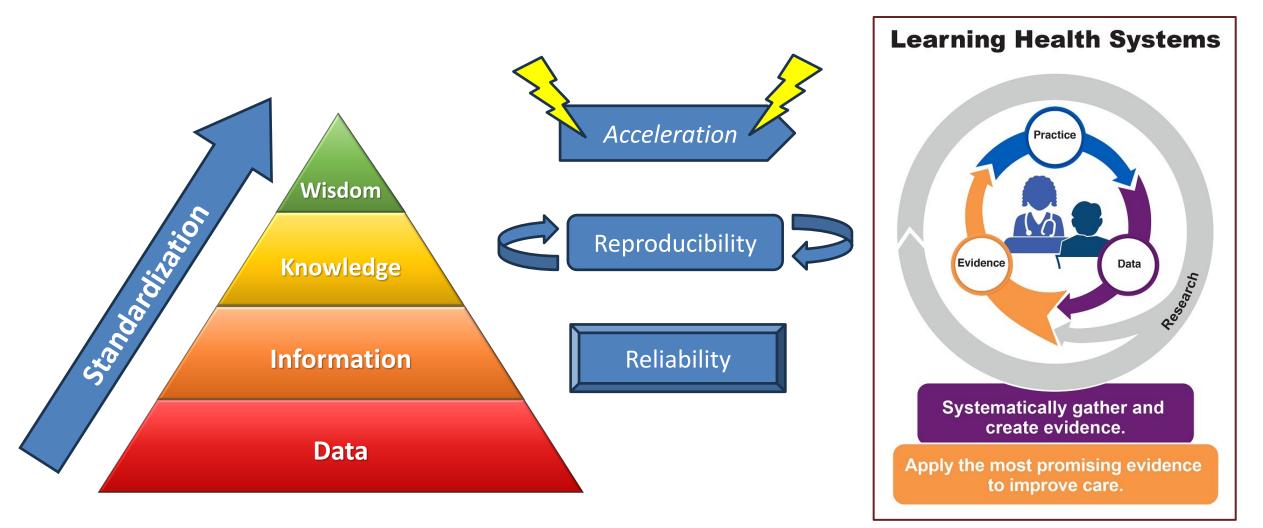
IterativeHardThresholding is an R package for performing L 0-based regressions using Cyclops.

OhdsiSharing

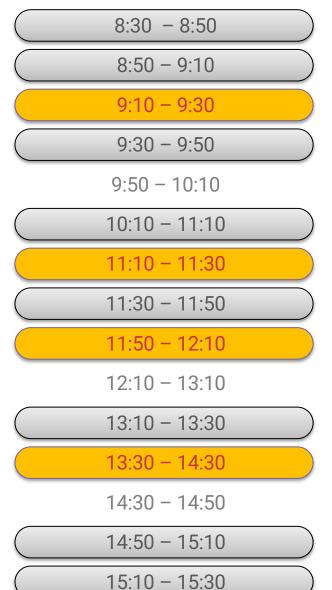
This is an R package for sharing data between OHDSI partners

Standardization of evidence generation \rightarrow acceleration, reproducibility, reliability











Value Prop 1: Standardization

How to frame an RWE research question

Value Prop 2 : Community, Open-source

BREAK

Panel: Stump the Experts

Atlas 1: Building Cohorts

Value Prop 3: Methods and Research

Atlas 2: Characterization

LUNCH

Value Prop 4: Reproducibility and Trust

Atlas 3: Group Exercise

BREAK

RWE Evidence at Scale

Take Home Message

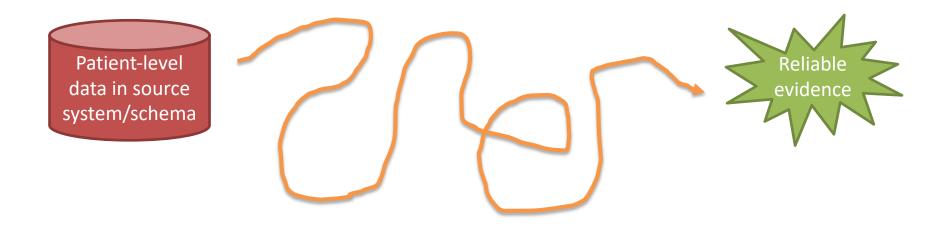


Evidence Generation

Large Scale Observational Research Preparation

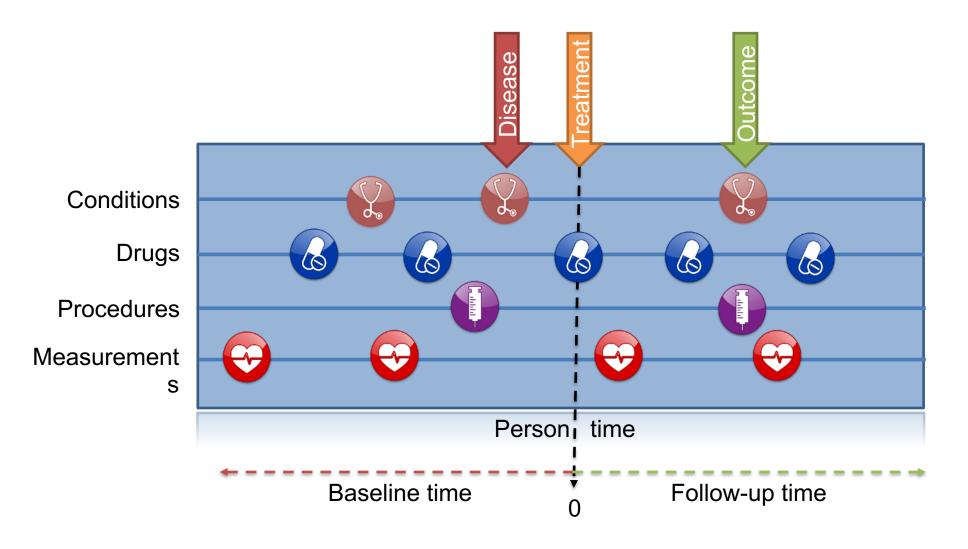


The journey to real-world evidence



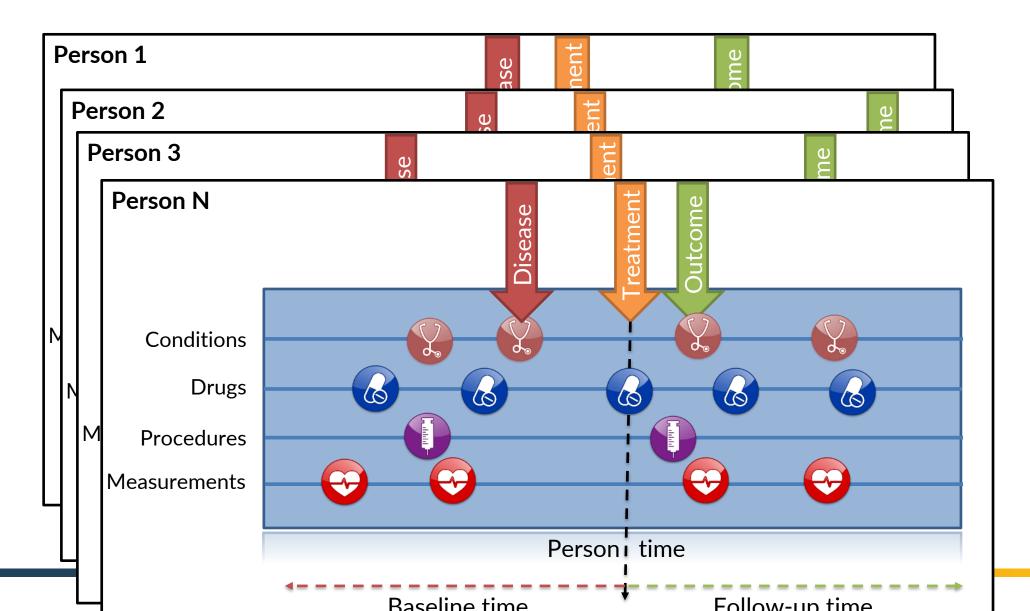


A Caricature of The Patient Journey



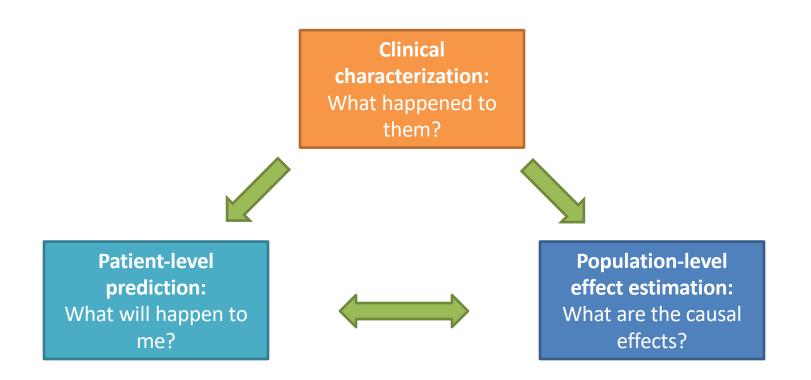


Each Observational Database Is Just an (Incomplete) Compilation of Patient Journeys





Complementary evidence to inform the patient journey



Analytic use case	Туре	Structure	Example
	Disease Natural History	Amongst patients who are diagnosed with <insert disease="" favorite="" your="">, what are the patient's characteristics from their medical history?</insert>	Amongst patients with rheumatoid arthritis , what are their demographics (age, gender), prior conditions, medications, and health service utilization behaviors?
Clinical characterization	Treatment utilization	Amongst patients who have <insert disease="" favorite="" your="">, which treatments were patients exposed to amongst <list of<br="">treatments for disease> and in which sequence?</list></insert>	Amongst patients with depression , which treatments were patients exposed to SSRI , SNRI , TCA , bupropion , esketamine and in which sequence?
	Outcome incidence	Amongst patients who are new users of <insert favorite<br="" your="">drug>, how many patients experienced <insert favorite<br="" your="">known adverse event from the drug profile> within <time horizon following exposure start>?</time </insert></insert>	Amongst patients who are new users of methylphenidate , how many patients experienced psychosis within 1 year of initiating treatment?
	Disease onset and progression	For a given patient who is diagnosed with <insert b="" favorite<="" your=""> disease>, what is the probability that they will go on to have <another complication="" disease="" or="" related=""></another> within <time b="" horizon<=""> from diagnosis>?</time></insert>	For a given patient who is newly diagnosed with atrial fibrillation , what is the probability that they will go onto to have ischemic stroke in next 3 years ?
Patient level prediction	Treatment response	For a given patient who is a new user of <insert favorite<br="" your="">chronically-used drug>, what is the probability that they will <insert desired="" effect=""> in <time window="">?</time></insert></insert>	For a given patient with T2DM who start on metformin , what is the probability that they will maintain HbA1C<6.5% after 3 years?
	Treatment safety	For a given patient who is a new user of <insert favorite<br="" your="">drug>, what is the probability that they will experience <insert adverse event > within <time exposure="" following="" horizon="">?</time></insert </insert>	For a given patients who is a new user of warfarin , what is the probability that they will have GI bleed in 1 year ?
Population-level	Safety surveillance	Does exposure to <insert drug="" favorite="" your=""> increase the risk of experiencing <insert adverse="" an="" event=""> within <time exposure="" following="" horizon="" start="">?</time></insert></insert>	Does exposure to ACE inhibitor increase the risk of experiencing Angioedema within 1 month after exposure start?
effect estimation	Comparative effectiveness	Does exposure to <insert drug="" favorite="" your=""> have a different risk of experiencing <insert (safety="" any="" benefit)="" or="" outcome=""> within <time exposure="" following="" horizon="" start="">, relative to <insert comparator="" treatment="" your="">?</insert></time></insert></insert>	Does exposure to ACE inhibitor have a different risk of experiencing acute myocardial infarction while on treatment , relative to thiazide diuretic ?



How does the rate of side effects / local problems (including secondary / palliative treatments needed) compare between treatments (open, laparoscopic, robot surgery, with or without lymph node dissection; brachytherapy, different forms of external beam radiation therapy), and which patient specific factors are associated with these adverse secondary endpoints?



 How does the rate of side effects / local problems (including secondary / palliative treatments needed) compare between
 treatments (open, laparoscopic, robot surgery, with or without lymph node
 dissection; brachytherapy, different forms
 of external beam radiation therapy) and

which patient specific factors are associated with these adverse secondary endpoints? <u>Characterization study: incidence rate</u>

Amongst patients with **prostate cancer receiving different treatments**, how many patients experienced **side effects/local problems** within <time horizon >?



 How does the rate of side effects / local problems (including secondary / palliative treatments needed) compare between
 treatments (open, laparoscopic, robot surgery, with or without lymph node
 dissection; brachytherapy, different forms
 of external beam radiation therapy) and

which patient specific factors are associated with these adverse secondary endpoints? <u>Population level estimation: comparative</u> <u>effectiveness</u>

Comparative effectiveness: Does exposure to **treatment A** have a different risk of experiencing **side effects/local problems** within <time horizon > , relative to **treatment B**?



- How does the rate of side effects / local problems (including secondary / palliative treatments needed) compare between treatments (open, laparoscopic, robot surgery, with or without lymph node dissection; brachytherapy, different forms of external beam radiation therapy), and which patient specific factors are associated with these adverse secondary endpoints?
- Characterization study: natural history

Amongst patients with **prostate cancer receiving different treatment A-Z**, what are the patient's characteristics from their medical history?



- RQ5. Which specific patient groups benefit ۲ most of upfront chemotherapy? What are the side effects and What is impact on quality of life in real-life practice of chemotherapy in this setting? the benefit of potentially toxic upfront chemotherapy appears to be highly individual. Other factors to predict who would benefit most are needed. the benefit of chemotherapy in the subgroup patients who have recurrence after primary treatment is not known.
- <u>Study</u>
- Target Cohorts:
- Comparator Cohorts:
- Outcome Cohorts:



Agenda





A collaborative open-science community transforming clinical research with real world evidence.

Paul Nagy, PhD, FSIIM Johns Hopkins University



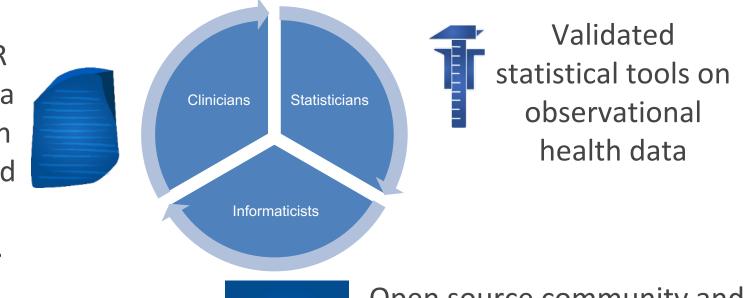


"You come for the data model, but you stay for the community"

Multi-disciplinary Innovation with Open Science at Scale



Translate EMR and claims data into a common data model tied to standard terminologies.



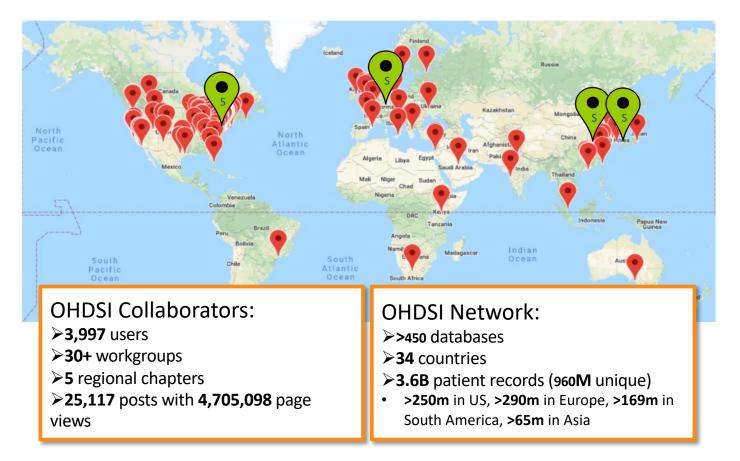


Open source community and research network with 900+ Million unique patients





OHDSI Is a Highly Active Global Community





OHDSI is a vibrant multi-speciality open science community

- Innovation
- Reproducibility
- •Community
- Collaboration
- •Openness
- •Beneficence

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

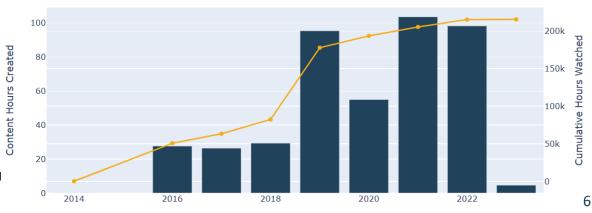


Education

- •Weekly Community Calls
- Phenotype Phebruary
- Save our Sysphus Challenge
- Open source developers conference



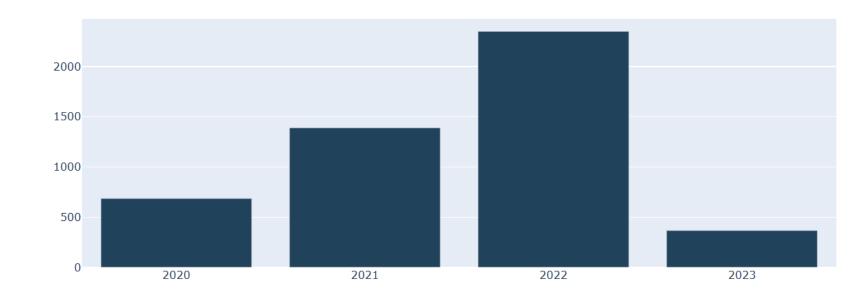
 Ehden Acdemy online learning mgmt. system. YouTube Analysis Book of OHDSI







Users by Year



- •Free online LMS based on Moodle
- •20+ self paced courses on OHDSI
- •https://academy.ehden.eu/



OHDSI Working Groups

		Workgroup
Core	Data model	Vocabulary
		Common Data Model
	Methods	Patient-Level Prediction
		Population-Level Estimation
	Tools	HADES
		Data Quality Dashboard
		Phenotype Development & Evaluation
		ATLAS/WebAPI
Domain	Data	Clinical Trials
	source	
		FHIR & OMOP
		Geographic Information Systems
		Medical Devices
		Medical Imaging
		Natural Language Processing
		Registry
		Vaccine Vocabulary
	Exposure	Health Equity
		Surgery and Perioperative Medicine

Domain	Disease	Oncology
		Psychiatry
		Eye Care
		Dentistry
Support	Regional	Africa Chapter
		Asia-Pacific Chapter
		Latin America Chapter
	Community segment	Early-Stage Researchers
		Open-Source Community
		Technical Advisory board
		Perseus Uses Group
		Databricks Users Group
		Healthcare Systems
	Broad	Education
		Steering Group

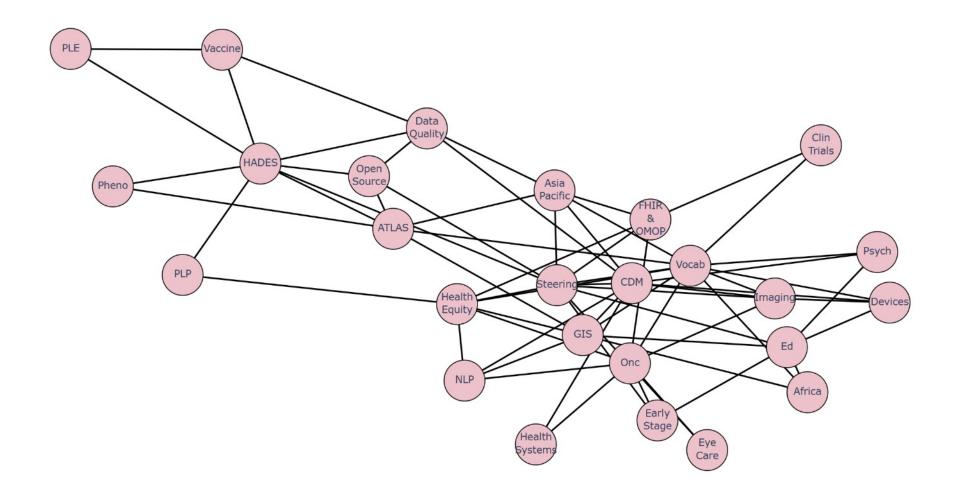


Working Groups

- •Any one can become an OHDSI member at no cost. OHDSI is an open inclusive community.
- •Any OHDSI member is welcome to join any working group.
- •Most working groups are 2x/Month, some are weekly.
- •Working group meetings are often recorded and if of educational nature will be uploaded to YouTube.



Working group collaboration





In Person Conferences

OHDSI Europe

OHDSI Asia Pacific





OHDSI Global Symposium





OHDSI Open Source Community

- 262 Repositories
- 30 M+ lines of code
- 681 Developers
- 31 organizations
- 47,672 commits
- 2,838 GitHub Forks
- 4,168 GitHub Stars
- 5,547 GitHub Subscribers

HADES 🏫 🍞 Package	s 🗹	alidation 🥟 Publi	cations Q	Support -		dy packages 👻	🔑 Develope
Package	Version	Maintainer(s)	Availability	Open issues	pull- requests	Build status	Coverage
Achilles	v1.7.2	Frank DeFalco	CRAN	29	3	() R check passing	P codecov 2
Andromeda	v0.6.3	Adam Black	CRAN	13	2	() R check passing	Condecov 89
BigKnn	v1.0.2	Martijn Schuemie	GitHub	0	0	C R check passing	
BrokenAdaptiveRidge	v1.0.0	Marc Suchard	CRAN	2	0	C R check passing	Codecov 9
Capr	v2.0.7	Martin Lavallee	GitHub	2	0	C R check passing	Codecov 8
Characterization	v0.1.2	Jenna Reps	GitHub	11	1	C R check failing	🗣 codecov 🔤 unkro
<u>CirceR</u>	v1.3.1	Chris Knoll	GitHub	3	1	O R check passing	Codecov 8
CohortDiagnostics	v3.2.4	Jamie Gilbert	GitHub	56	2	() R check passing	Codecov 8
CohortExplorer	v0.1.0	Gowtham Rao	CRAN	0	0	C R check passing	💎 codecov 🛛 10
CohortGenerator	v0.8.1	Anthony Sena	GitHub	19	2	(") R check passing	Codecov 9
CohortMethod	v5.1.0	Martijn Schuemie	GitHub	14	1	C R check failing	Codecov B
<u>Cyclops</u>	v3.3.1	Marc Suchard	CRAN	18	0	C R check failing	Codecov 8
DatabaseConnector	v6.2.4	Martijn Schuemie	CRAN	12	0	C R check passing	🗬 codecov 📴
DataQualityDashboard	v2.4.1	Katy Sadowksi	GitHub	43	7	C R check passing	Codecov 📴
DeepPatientLevelPrediction	v2.0.0	Egill Fridgeirsson	GitHub	18	1	C R check lailing	🖗 codecov 10
EmpiricalCalibration	v3.1.1	Martijn Schuemie	CRAN	1	0	() R check passing	Codecov B
EnsemblePatientLevelPrediction	v1.0.2	Jenna Reps	GitHub	5	0	() R check passing	👎 cadeone 🛛 unka
Eunomia	v1.0.2	Frank DeFalco	GitHub	10	1	C R check passing	Codecov Z
EvidenceSynthesis	v0.5.0	Martijn Schuemie	CRAN	3	0	() R check passing	Codecov 2
FeatureExtraction	v3.3.1	Anthony Sena	GitHub	44	6	C R check lailing	Codecov 9
Hydra	v0.4.0	Anthony Sena	GitHub	6	7	() R check lailing	Codecov 8
terativeHardThresholding	v1.0.2	Marc Suchard	CRAN	1	0	() R check passing	📿 codecov 🧧
MethodEvaluation	v2.3.0	Martijn Schuemie	GitHub	1	0	C R check passing	Codecov 3
OhdsiSharing	v0.2.2	Lee Evans	GitHub	0	1	C R check passing	Codecov 🚺
OhdsiShinyModules	v2.0.0	Jenna Reps	GitHub	108	2	C R check failing	Codecov Z
ParallelLogger	v3.3.0	Martijn Schuemie	CRAN	4	0	C R check passing	Codecov 8
PatientLevelPrediction	v6.3.6	Jenna Reps & Peter Rijnbeek	GitHub	47	0	R check failing	Codecov B
PhenotypeLibrary	v3.30.1	Gowtham Rao	GitHub	П	0	C R check passing	eodecov 10

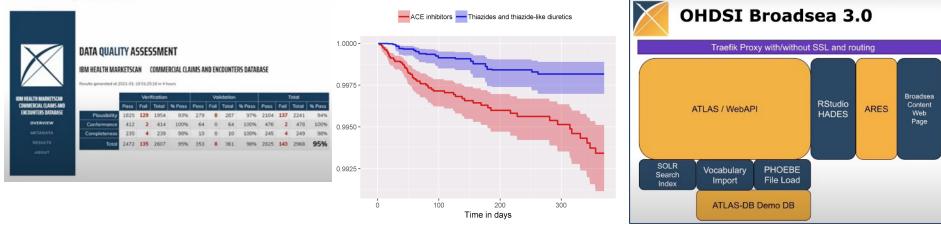
https://ohdsi.github.io/Hades/packageStatuses.html

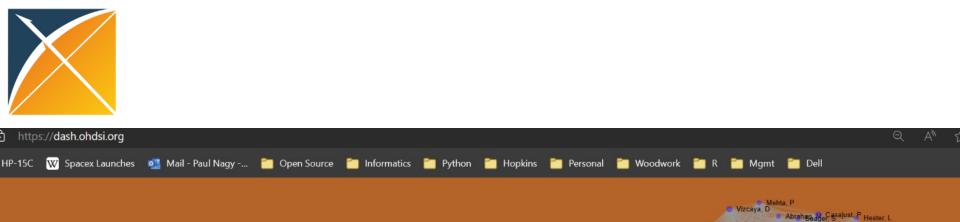


Data Science Applications

ATLAS		Achilles
🖶 Home	📸 Cohort #2459	
😑 Data Sources		Achilles President Data Services - III Reports -
Q Vocabulary		Dubboard
🛱 Concept Sets	Definition ② Concept Sets Generation Reporting Export	CDM Survey Population by Gender Age at Franc Diservation Secure name: (IM, (CAL)(V479)
P Cohort Definitions	Available CDM Sources	Number of persons (SG83M
Incidence Rates	Source Name Generation Status People Records Generated Generat	eration Duration
Profiles	▶ Generate ▼ COMPLETE 7,334 7,334	33.452s View Reports Constant Observator Person NIC Continues Observator NIC Continues Observator NIC Continues Observator
Estimation	4:57:05	33.4525 View Reports Constant Ty More
Prediction	Inclusion Report Cohort Features	
🖬 Jobs	/ /	
Configuration	Inclusion Report for Match Rate Matches Total	Switch to intersect view
🗩 Feedback	Summary Statistics: 96.39% 85,602 88,812	
	Inclusion Rule N % Remain % Diff	
	1. 85,602 96.39% 3.61%	

Data Quality Dashboard



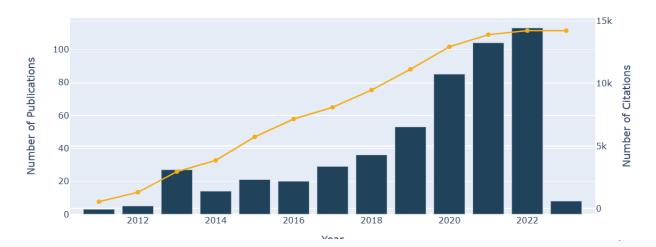






Publications

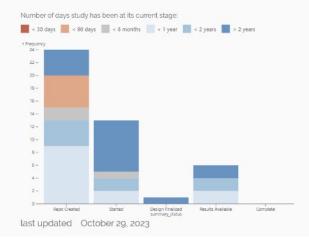
OHDSI Publications & Cumulative Citations



	Journal	Creation Date	Authors
Methods for drug safety signal detection in longitudinal observational databases: LGPS and LEOPARD.	Pharmacoepidemiology and drug safety	2010/10/15 06:00	Schuemie. Martijn J
Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership.	Annals of internal medicine	2010/11/03 06:00	Stang, Paul E Ryan. Patrick B Racoosin, Judith A Overhage, J Marc Hartzema, Abraham G Reich, Christian Welebob, Emily Scarnecchia, Thomas Woodcock, Janet
Validation of a common data model for active safety surveillance research. 🔀	Journal of the American Medical Informatics Association : JAMIA	2011/11/01 06:00	Overhage, J Marc Ryan, Patrick B Reich, Christian G Hartzema, Abraham G Stang, Paul E
Mini-Sentinel's systematic reviews of validated methods for identifying health outcomes using administrative and claims data: methods and lessons learned.	Pharmacoepidemiology and drug safety	2012/01/21 06:00	Carnahan, Ryan M Moores, Kevin G
Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases.	Journal of biomedical informatics	2012/06/12 06:00	Reich, Christian Ryan, Patrick B Stang, Paul E Rocca, Mitra



Creating Evidence



/dash.ohdsi.org/ph											
W Spacex Launches	🥶 Mail - Paul Nagy	🛅 Open Source	🛅 Informatics	🛅 Python 🛅 He	pkins 🎦 Persona	📔 Woodwork	🛅 R	🛅 Mgmt	🛅 Dell		
				Publicatio	ons Data Net	work Open	Source	L Learn	More		

STUDIES STUDY LEADS DATA PARTNERS

rd

Phenotypes

Cohorts

		Q,
	Status	Updated
Naussa or Vomiting 🖸	Pending peer review	2023-09-25
Alzheimer's disease [2]	Pending peer review	2023-09-25
Nauson3 🖸	Pending	2023-09-20
Peripheral edema [2]	Pending	2023-10-09
Photosensitivity 🔀	Pending	2023-09-24
Renal cancer	Pending	2023-09-24
Thyroid turner 🖸	Pending	2023-10-05
Venous thromboembolism 🖸	Pending	2023-09-24
Vomiling symptoms	Pending	2023-09-24

		Time at Stage	Last Update	Protocol	Results
AlcoholicLiverDisease	Started	1053 days	2021-09-13 14:33:29	因	
EumaeusCovid1gVaccines	Started	1032 days	2021-07-06 10:08:56		
AesiIncidenceCorrection	Repo Created	236 days	2023-03-19 19:14:58	五	Ń
DoacsWarfarinSub	Results Available	1060 days	2020-12-15 15:01:54	因	Í
DeepLearningComparison	Repo Created	358 days	2023-01-24 15:17:52		
FluoroquinoloneAorticAneurysm	Repo Created	226 days	2023-04-18 14:41:01		
SosTesti	Repo Created	235 days	2023-03-28 15:51:15		
DeconfounderEvaluation	Repo Created	1075 days	2020-11-30 14:03:20		
MultipleSclerosisBiologicsPml	Repo Created	226 days	2023-04-18 14:40:30		
IbdCharacterization	Repo Created	886 days	2022-07-31 10:25:00	因	Ń
NephrologyHealthEquity	Started	365 days	2022-11-16 20:58:28	因	



https://ohdsi.org

: 👿 Spacex Launches 📴 Mail - Paul Nagy 🎦 Open Source 📁 I	Informatics 🛅 Python 🞽 Hopkins 🚰 Personal 🎽 Woodwork 🎽 R 🎽 Mgmt 🎽 Dell OHDSI Home Forums Wiki Github
	OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS
Who We Are	Standards Software Tools - Network Studies - Community Forums - Education - New To OHDSI? - Workgroups - 2023 'Our Journey' Annual Report Community Dashboards This Week In OHDSI
OHDSI Publications Support & Support	s Learn About Our Workgroups posium ~ Github YouTube Twitter LinkedIn Newsletters ~
	Join Our Workgroups Workgroup Call Schedule Vorkgroup S
	Best Practices in MS Teams
evidence that promote	on is to improve health by empowering a community to collaboratively generate the s better health decisions and better care. We work towards that goal in the areas of dological research, open-source analytics development, and clinical applications.

Agenda





Stump the Experts

Panel Session



Panelists





Mui Van Zandt VP/Global Head Data Strategy, Access & Enablement - GM Inteliquet IQVIA



Paul Nagy

Program Director for Graduate Training in Biomedical Informatics and Data Science, Deputy Director of the Johns Hopkins Medicine Technology Innovation Center Johns Hopkins University



Christian Reich

Professor of PracticeProfessor of Practice; Northeastern University, CEO; Odysseus Data Services,





45 min – moderated questions

15 min – Audience questions

AMIA 2023 Annual Symposium | amia.org

Challenge our OHDSI panelists at AMIA!



Submit your most intriguing questions and be a part of our 'Stump the Experts' session!



Join by Web

PollEv.com/clt3

Join by Text

Send clt3 and your message to 37607

Agenda







OHDSI RWE Revolution:

Igniting Data Modernization with Harmonized Standards

for Cutting-Edge Health Research

11-Nov-2023



Demo 1 Atlas 1: Building cohorts



Demo 2

Atlas 2: Characterization and visualization



Hands-on Session

Atlas 3: Group Exercise

Agenda







OHDSI Methods and Research

OHDSI RWE Revolution: Igniting Data Modernization with Harmonized Standards for Cutting-Edge Health Research

Gowtham A Rao

Johnson and Johnson Connect with me <u>rao@ohdsi.org</u> #AMIA2023





I disclose the following relevant relationship with commercial interests:

- Employee of Johnson and Johnson
- Spouse is employee of Johnson and Johnson

Much of the slides and content borrowed with permission from other OHDSI collaborators

• Martijn Schuemie, Patrick Ryan, Marc Suchard, Anthony Sena etc.

Standard Framework for Research Questions: TCIO-TAR inputs



What are the standardized inputs?

Target (T): The exposure of interest

Comparator (C): A suitable comparator

Indication (I): Ensure prior membership in an underlying disease cohort (optional)

- **Outcome (O):** Includes primary and secondary health status of interest either from an efficacy or safety perspective
- Time at Risk (TAR): The a priori determined period of time upon which the outcome is assessed

Example Research: TCIO-TAR for Diabetes Mellitus

T: SGLT2i

C: GLP-1RA

I: T2D

O: MACE, HHF, DKA, genital infections, fractures, LLA, AKI, UTI, mortality

TAR: On treatment

Age: >66

Characterization: Differences in baseline characteristics between T and C.

Estimation: Difference in risk between T and C for the O in the TAR.

Prediction: Occurrence of O among T within TAR

•

826

Comparative Effectiveness and Safety of Sodium–Glucose Cotransporter 2 Inhibitors Versus Glucagon-Like Peptide 1 Receptor Agonists in Older Adults

OBJECTIVE

Both sodium–glucose cotransporter 2 inhibitors (SGLT2i) and glucagon-like peptide 1 receptor agonists (GLP-1RA) demonstrated cardiovascular benefits in randomized controlled trials of patients with type 2 diabetes (T2D) generally <65 years old and mostly with cardiovascular disease. We aimed to evaluate the comparative effectiveness and safety of SGLT2i and GLP-1RA among real-world older adults.

RESEARCH DESIGN AND METHODS

Using Medicare data (April 2013–December 2016), we identified 90,094 propensity score–matched (1:1) T2D patients ≥66 years old initiating SGLT2i or GLP-1RA. Primary outcomes were major adverse cardiovascular events (MACE) (i.e., myocardial infarction, stroke, or cardiovascular death) and hospitalization for heart failure (HHF). Other outcomes included diabetic ketoacidosis (DKA), genital infections, fractures, lower-limb amputations (LLA), acute kidney injury (AKI), severe urinary tract infections, and overall mortality. We estimated hazard ratios (HRs) and rate differences (RDs) per 1,000 person-years, controlling for 140 baseline covariates.



Diabetes Care Volume 44, March 2021

Example Research: TCIO-TAR for Hormone Replacement Therapy



T: CE/BZA

C: EP

I: none

O: endometrial cancer, endometrial hyperplasia, and breast cancer (and others in the methods section)

TAR: On treatment

Characterization: Differences in baseline characteristics between T and C.

Estimation: Difference in risk between T and C for the O in the TAR.

Prediction: Occurrence of O among T within TAR

ORIGINAL STUDY

Comparative safety of conjugated estrogens/bazedoxifene versus estrogen/progestin combination hormone therapy among women in the United States: a multidatabase cohort study

Hoffman, Sarah R. MS, MPH, PhD¹; Governor, Samuel MD, MPH¹; Daniels, Kimberly PhD, MS¹; Seals, Ryan M. MPH, ScD²; Ziyadeh, Najat J. MA, MPH²; Wang, Florence T. ScD²; Dai, Dingwei MD, PhD³; Mcmahill-Walraven, Cheryl N. MSW, PhD³; Shuminski, Patty AS³; Frajzyngier, Vera PhD⁴; Zhou, Xiaofeng PhD⁴; Shen, Rongjun MS⁴; Garg, Renu K. PhD, MPH⁴; Fournakis, Nicole MPH¹; Lanes, Stephan PhD¹; Beachler, Daniel C. MHS, PhD¹

Author Information⊗

Menopause 30(8):p 824-830, August 2023. | DOI: 10.1097/GME.00000000002217 😁

OPEN SDC

Metrics

Abstract In Brief

Objective

To assess the risk of select safety outcomes including endometrial cancer, endometrial hyperplasia, and breast cancer among women using conjugated estrogens/bazedoxifene (CE/BZA) as compared with estrogen/progestin combination hormone therapy (EP).

Methods

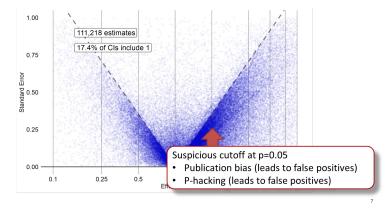
We conducted a new-user cohort study in five US healthcare claims databases representing more than 92 million women. We included CE/BZA or EP new users from May 1, 2014, to August 30, 2019. EP users were propensity score (PS) matched to users of CE/BZA. Incidence of endometrial cancer, endometrial hyperplasia, breast cancer, and eight additional cancer and cardiovascular outcomes were ascertained using claims-based algorithms. Rate ratios (RR) and differences pooled across databases were estimated using random-effects models.



We can answer a large set of questions using TCIO framework – but are our results trustable?

Problem: Long-standing issues of lack of TRUST in real-world evidence to guide clinical practice.

- Observational study bias
- Publication bias
- P-hacking
- Misleading estimates due to study design and analytical choices



Published observational study results

How do we earn and prevent erosion of TRUST in our science?



Three ideas

2. Objective Diagnostics

Phenotype development and evaluation

Fail

Engineering open science systems that build trust into the real-world evidence generation and dissemination process

Distributed data network, standardized to common data model

Network coordinatior

Analysis reliability evaluation

3. Standardized software



Open-source software



depend on the estimated effects. All generated evidence is disseminated at once. All million with rublication has and enhances transparent 3. LEGEND will generate evidence using a prespecified analysis design. All analyses, including the research questions that will

METHODS RESEARCH

4. LEGEND will generate evidence by consistently applying a systematic process across all research questions. This principle precludes modification of analyses to obtain a desired answer to any specific question. This does not imply a simple one-size-fits-all process, rather that the logic for modifying an analysis for specific research questions should be explicated and applied systematically. Aim: Avoids P hacking and allows for the evaluation of the operating characteristics of this process (Principle 6)

Principles of Large-scale Evidence Generation and

Evaluation across a Network of Databases (LEGEND)

Martijn J. Schuemie (3^{1,2}, Patrick B. Ryan^{1,3}, Nicole Pratt⁴, RuiJun Chen (3^{1,5}, Seng Chan You⁶, Harlan M. Krumholz⁷, David Madigan⁶, George Hripcsak^{2,6}, and

1. LEGEND Principles

LEGEND in Principle

LEGEND (Large-scale Evidence Generation and Evaluation across a Network of Databases) applies high-level analytics to perform observational research on hundreds

of millions of patient records within OHDSI's international database network. LEGEND is based on 10 guiding principles that were published in JAMIA (August, 2020)

Perspective

Marc A. Suchard^{2,1}

and are listed below

1. LEGEND will generate evidence at a large scale. Instead of answering a single question

at a time (on the effect of 1 treatment on 1 outcome).

LEGEND answers large sets of related questions at once (eg, the effects of many treatments for a disease

on many outcomes). Aim: Avoids publication bias,

achieves commehensiveness of results, and allows for

an evaluation of the overall coherence and consistency of the generated evidence.

2. Dissemination of the evidence will not

be answered, will be decided prior to analysis execution. Aim: Avoids P hacking

5. LEGEND will generate evidence using best practices. LEGEND answers each question using current best practices, including advanced methods to address confounding, such as propensity scores. Specifically, we will not employ suboptimal methods (in terms of bias) to achieve better computational efficiency. Aim: Minimizes bias

6. LEGEND will include empirical evaluation through the use of control questions. Every LEGEND study includes control questions. Control questions are questions where the answer is known. These allow for measuring the operating characteristics of our systematic process including residual bias. We subsequently account for this observed residual bias in our P values, effect estimates and confidence intervals using empirical calibration. [7.8] Aim: Enhances transparency on the uncertainty due to residual bias

7. LEGEND will generate evidence using open-source software that is freely available to all. The analysis software is com to review and evaluation and is available for replicating analyses down to the smallest detail Aim; Enhances transparency and allows replication

8. LEGEND will not be used to evaluate new methods. Even though the same intrastructure used in LEGEND may also be used to evaluate new causal inference methods, generating clinical evidence should not be performed at the same time as method evaluation. This is a corollar of Principle 5, since a new method that still requires evaluation cannot already be best practice. Also, generating evidence with unproven methods can harmore the interpretability of the clinical results. Note that I EGEND does evaluate how well the methods it uses perform in the specific context of the questions and data used in a LEGEND study (Principle 6). Alm: Avoids bias and improves interpretability.

9. LEGEND will generate evidence across a network of multiple databases. Multiple heterogeneous databases (different data capture processes, health care systems, and populations) will be used to generate the evidence to allow an assessment of the replicability of findings across sites. AIm: Enhances generalizability and uncovers potential between-site heterogeneity.

10. LEGEND will maintain data confidentiality; patient-level data will not be shared between sites in the network Not sharing data will ensure patient privacy, and comply with local data governance rules. Aim: Privacy

#JoinTheJourney	47	OHDSI.org

System characteristics:

Data quality evaluation

- Standardized procedures with defined inputs and outputs
- Analysis packages implementing scientific best practices
- consistently applied across all data partners, generating consistent output for network synthesis
- Reproducible outputs generated by open-source analysis libraries developed and validated with verifiable unit-test coverage
- Pre-specified and objective decision thresholds for go/no go criteria
- Measurable operating characteristics of system performance





Journal of the American Medical Informatics Association, 2020, Vol. 27, No. 8

1333

Table 1: Guiding principles of the Large-scale Evidence Generation and Evaluation across a Network of Databases (LEGEND) initiative.

1 LEGEND will generate evidence at a large scale.

Instead of answering a single question at a time (eg, the effect of 1 treatment on 1 outcome), LEGEND answers large sets of related questions at once (eg, the effects of many treatments for a disease on many outcomes).

- Aim: Avoids publication bias, achieves comprehensiveness of results, and allows for an evaluation of the overall coherence and consistency of the generated evidence.
- 2 Dissemination of the evidence will not depend on the estimated effects.

All generated evidence is disseminated at once.

Aim: Avoids publication bias and enhances transparency.

3 LEGEND will generate evidence using a prespecified analysis design.

All analyses, including the research questions that will be answered, will be decided prior to analysis execution. Aim: Avoids P hacking.

4 LEGEND will generate evidence by consistently applying a systematic process across all research questions.

This principle precludes modification of analyses to obtain a desired answer to any specific question. This does not imply a simple one-size-fits-all process, rather that the logic for modifying an analysis for specific research questions should be explicated and applied systematically. A im: Avoids P hacking and allows for the evaluation of the operating characteristics of this process (Principle 6).

5 LEGEND will generate evidence using best practices.

LEGEND answers each question using current best practices, including advanced methods to address confounding, such as propensity scores. Specifically, we will not employ suboptimal methods (in terms of bias) to achieve better computational efficiency. Aim: Minimizes bias.

6 LEGEND will include empirical evaluation through the use of control questions.

Every LEGEND study includes control questions. Control questions are questions where the answer is known. These allow for measuring the operating characteristics of our systematic process, including residual bias. We subsequently account for this observed residual bias in our P values, effect estimates, and confidence intervals using empirical calibration.[7,8]

Aim: Enhances transparency on the uncertainty due to residual bias.

- 7 LEGEND will generate evidence using open-source software that is freely available to all. The analysis software is open to review and evaluation, and is available for replicating analyses down to the smallest detail. Aim: Enhances transarence and allows replication.
- 8 LEGEND will not be used to evaluate new methods.

Even though the same infrastructure used in LEGEND may also be used to evaluate new causal inference methods, generating clinical evidence should not be performed at the same time as method evaluation. This is a corollary of Principle 5, since a new method that still requires evaluation cannot already be best practice. Also, generating evidence with unproven methods can hamper the interpretability of the clinical results. Note that LEGEND does evaluate how well the methods it uses perform in the specific context of the questions and data used in a LEGEND study (Principle 6).

Aim: Avoids bias and improves interpretability.

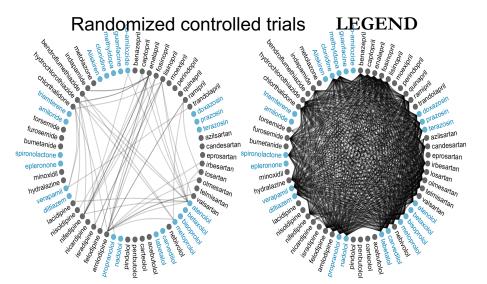
9 LEGEND will generate evidence across a network of multiple databases.

Multiple heterogeneous databases (different data capture processes, health-care systems, and populations) will be used to generate the evidence to allow an assessment of the replicability of findings across sites.

Aim: Enhances generalizability and uncovers potential between-site heterogeneity.

10 LEGEND will maintain data confidentiality; patient-level data will not be shared between sites in the network. Not sharing data will ensure patient privacy, and comply with local data governance rules. Aim: Privacy.

Note: LEGEND: Large-scale Evidence Generation and Evaluation across a Network of Databases.





Published observational study results Journal of the American Medical Informatics Association, 2020, Vol. 27, No. 8 1333 Idea Perform study Submit paper Publication! 1.00 111,218 estimates Table 1: Guiding principles of the Large-scale Evidence Generation and Evaluation across a Network of Databases (LEGEND) initiative. 17.4% of CIs include 1 1 LEGEND will generate evidence at a large scale. Instead of answering a single question at a time (eg, the effect of 1 treatment on 1 outcome), LEGEND answers large sets of related questions at once (eg. the effects of many treatments for a disease on many outcomes). Aim: Avoids publication bias, achieves comprehensiveness of results, and allows for an evaluation of the overall coherence and consistency of the generated evidence. uspicious cutoff at p=0.05 2 Dissemination of the evidence will not depend on the estimated effects. Publication bias (leads to false positives) All generated evidence is disseminated at once. 0.1 P-hacking (leads to false positives) Aim: Avoids publication bias and enhances transparency. 3 LEGEND will generate evidence using a prespecified analysis design. All analyses, including the research questions that will be answered, will be decided prior to analysis execution. Aim: Avoids P hacking. 4 LEGEND will generate evidence by consistently applying a systematic process across all research questions. This principle precludes modification of analyses to obtain a desired answer to any specific question. This does not imply a simple one-size-fits-all process, rather that the logic for modifying an analysis for specific research questions should be explicated and applied systematically. Aim: Avoids P hacking and allows for the evaluation of the operating characteristics of this process (Principle 6). LEGEND will generate evidence using best practices. 5 LEGEND answers each question using current best practices, including advanced methods to address confounding, such as propensity scores, Specifically, we will not employ suboptimal methods (in terms of bias) to achieve better computational efficiency. Aim: Minimizes bias. 6 LEGEND will include empirical evaluation through the use of control questions. Every LEGEND study includes control questions. Control questions are questions where the answer is known. These allow for measuring the op-P-hacking erating characteristics of our systematic process, including residual bias. We subsequently account for this observed residual bias in our P values, effect estimates, and confidence intervals using empirical calibration.[7,8] Publication! Aim: Enhances transparency on the uncertainty due to residual bias. Idea Perform study Submit paper 7 LEGEND will generate evidence using open-source software that is freely available to all. The analysis software is open to review and evaluation, and is available for replicating analyses down to the smallest detail. Aim: Enhances transparency and allows replication. 8 LEGEND will not be used to evaluate new methods. Even though the same infrastructure used in LEGEND may also be used to evaluate new causal inference methods, generating clinical evidence should not be performed at the same time as method evaluation. This is a corollary of Principle 5, since a new method that still requires evaluation cannot already be best practice. Also, generating evidence with unproven methods can hamper the interpretability of the clinical results. Note that LEGEND does evaluate how well the methods it uses perform in the specific context of the questions and data used in a LEGEND study (Principle 6). Aim: Avoids bias and improves interpretability. 9 LEGEND will generate evidence across a network of multiple databases. Multiple heterogeneous databases (different data capture processes, health-care systems, and populations) will be used to generate the evidence to allow an assessment of the replicability of findings across sites. Aim: Enhances generalizability and uncovers potential between-site heterogeneity. 10 LEGEND will maintain data confidentiality: patient-level data will not be shared between sites in the network. Not sharing data will ensure patient privacy, and comply with local data governance rules. Aim: Privacy.

Note: LEGEND: Large-scale Evidence Generation and Evaluation across a Network of Databases.

Publication bias





Journal of the American Medical Informatics Association, 2020, Vol. 27, No. 8

1333

Table 1: Guiding principles of the Large-scale Evidence Generation and Evaluation across a Network of Databases (LEGEND) initiative.

1 LEGEND will generate evidence at a large scale.

Instead of answering a single question at a time (eg, the effect of 1 treatment on 1 outcome), LEGEND answers large sets of related questions at once (eg, the effects of many treatments for a disease on many outcomes).

Aim: Avoids publication bias, achieves comprehensiveness of results, and allows for an evaluation of the overall coherence and consistency of the generated evidence.

2 Dissemination of the evidence will not depend on the estimated effects.

All generated evidence is disseminated at once.

- Aim: Avoids publication bias and enhances transparency.
- 3 LEGEND will generate evidence using a prespecified analysis design.

All analyses, including the research questions that will be answered, will be decided prior to analysis execution. Aim: Avoids P hacking.

4 LEGEND will generate evidence by consistently applying a systematic process across all research questions.

This principle precludes modification of analyses to obtain a desired answer to any specific question. This does not imply a simple one-size-fits-all process, rather that the logic for modifying an analysis for specific research questions should be explicated and applied systematically. A im: Avoids P hacking and allows for the evaluation of the operating characteristics of this process (Principle 6).

5 LEGEND will generate evidence using best practices.

LEGEND answers each question using current best practices, including advanced methods to address confounding, such as propensity scores. Specifically, we will not employ suboptimal methods (in terms of bias) to achieve better computational efficiency. Aim: Minimizes bias.

6 LEGEND will include empirical evaluation through the use of control questions.

Every LEGEND study includes control questions. Control questions are questions where the answer is known. These allow for measuring the operating characteristics of our systematic process, including residual bias. We subsequently account for this observed residual bias in our P values, effect estimates, and confidence intervals using empirical calibration [7,8]

Aim: Enhances transparency on the uncertainty due to residual bias.

- 7 LEGEND will generate evidence using open-source software that is freely available to all. The analysis software is open to review and evaluation, and is available for replicating analyses down to the smallest detail. Aim: Enhances transarence and allows replication.
- 8 LEGEND will not be used to evaluate new methods.

Even though the same infrastructure used in LEGEND may also be used to evaluate new causal inference methods, generating clinical evidence should not be performed at the same time as method evaluation. This is a corollary of Principle 5, since a new method that still requires evaluation cannot already be best practice. Also, generating evidence with unproven methods can hamper the interpretability of the clinical results. Note that LEGEND does evaluate how well the methods it uses perform in the specific context of the questions and data used in a LEGEND study (Principle 6).

Aim: Avoids bias and improves interpretability.

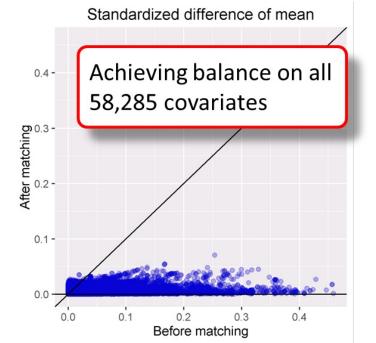
9 LEGEND will generate evidence across a network of multiple databases.

Multiple heterogeneous databases (different data capture processes, health-care systems, and populations) will be used to generate the evidence to allow an assessment of the replicability of findings across sites.

Aim: Enhances generalizability and uncovers potential between-site heterogeneity.

10 LEGEND will maintain data confidentiality; patient-level data will not be shared between sites in the network. Not sharing data will ensure patient privacy, and comply with local data governance rules. Aim: Privacy.

Note: LEGEND: Large-scale Evidence Generation and Evaluation across a Network of Databases.



Journal of the American Medical Informatics Association, 2020, Vol. 27, No. 8

1333

Table 1: Guiding principles of the Large-scale Evidence Generation and Evaluation across a Network of Databases (LEGEND) initiative.

1 LEGEND will generate evidence at a large scale.

Instead of answering a single question at a time (eg, the effect of 1 treatment on 1 outcome), LEGEND answers large sets of related questions at once (eg, the effects of many treatments for a disease on many outcomes).

Aim: Avoids publication bias, achieves comprehensiveness of results, and allows for an evaluation of the overall coherence and consistency of the generated evidence.

2 Dissemination of the evidence will not depend on the estimated effects.

All generated evidence is disseminated at once.

Aim: Avoids publication bias and enhances transparency.

3 LEGEND will generate evidence using a prespecified analysis design.

All analyses, including the research questions that will be answered, will be decided prior to analysis execution. Aim: Avoids P hacking.

4 LEGEND will generate evidence by consistently applying a systematic process across all research questions.

This principle precludes modification of analyses to obtain a desired answer to any specific question. This does not imply a simple one-size-fits-all process, rather that the logic for modifying an analysis for specific research questions should be explicated and applied systematically. A im: Avoids P hacking and allows for the evaluation of the operating characteristics of this process (Principle 6).

5 LEGEND will generate evidence using best practices.

LEGEND answers each question using current best practices, including advanced methods to address confounding, such as propensity scores. Specifically, we will not employ suboptimal methods (in terms of bias) to achieve better computational efficiency.

Aim: Minimizes bias.

6 LEGEND will include empirical evaluation through the use of control questions.

Every LEGEND study includes control questions. Control questions are questions where the answer is known. These allow for measuring the operating characteristics of our systematic process, including residual bias. We subsequently account for this observed residual bias in our P values, effect estimates, and confidence intervals using empirical calibration [7,8]

Aim: Enhances transparency on the uncertainty due to residual bias.

7 LEGEND will generate evidence using open-source software that is freely available to all.

The analysis software is open to review and evaluation, and is available for replicating analyses down to the smallest detail. Aim: Enhances transparency and allows replication.

8 LEGEND will not be used to evaluate new methods.

Even though the same infrastructure used in LEGEND may also be used to evaluate new causal inference methods, generating clinical evidence should not be performed at the same time as method evaluation. This is a corollary of Principle 5, since a new method that still requires evaluation cannot already be best practice. Also, generating evidence with unproven methods can hamper the interpretability of the clinical results. Note that LEGEND does evaluate how well the methods it uses perform in the specific context of the questions and data used in a LEGEND study (Principle 6).

Aim: Avoids bias and improves interpretability.

9 LEGEND will generate evidence across a network of multiple databases.

Multiple heterogeneous databases (different data capture processes, health-care systems, and populations) will be used to generate the evidence to allow an assessment of the replicability of findings across sites.

Aim: Enhances generalizability and uncovers potential between-site heterogeneity.

10 LEGEND will maintain data confidentiality; patient-level data will not be shared between sites in the network. Not sharing data will ensure patient privacy, and comply with local data governance rules. Aim: Privacy.

Note: LEGEND: Large-scale Evidence Generation and Evaluation across a Network of Databases.



Measuring residual bias



Journal of the American Medical Informatics Association, 2020, Vol. 27, No. 8

1333

Table 1: Guiding principles of the Large-scale Evidence Generation and Evaluation across a Network of Databases (LEGEND) initiative.

1 LEGEND will generate evidence at a large scale.

Instead of answering a single question at a time (eg, the effect of 1 treatment on 1 outcome), LEGEND answers large sets of related questions at once (eg, the effects of many treatments for a disease on many outcomes).

Aim: Avoids publication bias, achieves comprehensiveness of results, and allows for an evaluation of the overall coherence and consistency of the generated evidence.

2 Dissemination of the evidence will not depend on the estimated effects.

All generated evidence is disseminated at once.

Aim: Avoids publication bias and enhances transparency.

3 LEGEND will generate evidence using a prespecified analysis design.

All analyses, including the research questions that will be answered, will be decided prior to analysis execution. Aim: Avoids P hacking.

4 LEGEND will generate evidence by consistently applying a systematic process across all research questions.

This principle precludes modification of analyses to obtain a desired answer to any specific question. This does not imply a simple one-size-fits-all process, rather that the logic for modifying an analysis for specific research questions should be explicated and applied systematically. A im: Avoids P hacking and allows for the evaluation of the operating characteristics of this process (Principle 6).

5 LEGEND will generate evidence using best practices.

LEGEND answers each question using current best practices, including advanced methods to address confounding, such as propensity scores. Specifically, we will not employ suboptimal methods (in terms of bias) to achieve better computational efficiency. Aim: Minimizes bias.

6 LEGEND will include empirical evaluation through the use of control questions.

Every LEGEND study includes control questions. Control questions are questions where the answer is known. These allow for measuring the operating characteristics of our systematic process, including residual bias. We subsequently account for this observed residual bias in our P values, effect estimates, and confidence intervals using empirical calibration [7,8]

Aim: Enhances transparency on the uncertainty due to residual bias.

7 LEGEND will generate evidence using open-source software that is freely available to all.

The analysis software is open to review and evaluation, and is available for replicating analyses down to the smallest detail. Aim: Enhances transparency and allows replication.

8 LEGEND will not be used to evaluate new methods.

Even though the same infrastructure used in LEGEND may also be used to evaluate new causal inference methods, generating clinical evidence should not be performed at the same time as method evaluation. This is a corollary of Principle 5, since a new method that still requires evaluation cannot already be best practice. Also, generating evidence with unproven methods can hamper the interpretability of the clinical results. Note that LEGEND does evaluate how well the methods it uses perform in the specific context of the questions and data used in a LEGEND study (Principle 6).

Aim: Avoids bias and improves interpretability.

9 LEGEND will generate evidence across a network of multiple databases.

Multiple heterogeneous databases (different data capture processes, health-care systems, and populations) will be used to generate the evidence to allow an assessment of the replicability of findings across sites.

Aim: Enhances generalizability and uncovers potential between-site heterogeneity.

10 LEGEND will maintain data confidentiality; patient-level data will not be shared between sites in the network. Not sharing data will ensure patient privacy, and comply with local data governance rules. Aim: Privacy.

Note: LEGEND: Large-scale Evidence Generation and Evaluation across a Network of Databases.



Open-source software



Journal of the American Medical Informatics Association, 2020, Vol. 27, No. 8

1333

Table 1: Guiding principles of the Large-scale Evidence Generation and Evaluation across a Network of Databases (LEGEND) initiative

1 LEGEND will generate evidence at a large scale.

Instead of answering a single question at a time (eg, the effect of 1 treatment on 1 outcome), LEGEND answers large sets of related questions at once (eg, the effects of many treatments for a disease on many outcomes).

Aim: Avoids publication bias, achieves comprehensiveness of results, and allows for an evaluation of the overall coherence and consistency of the generated evidence.

2 Dissemination of the evidence will not depend on the estimated effects.

All generated evidence is disseminated at once.

Aim: Avoids publication bias and enhances transparency.

3 LEGEND will generate evidence using a prespecified analysis design.

All analyses, including the research questions that will be answered, will be decided prior to analysis execution. Aim: Avoids P hacking.

4 LEGEND will generate evidence by consistently applying a systematic process across all research questions.

This principle precludes modification of analyses to obtain a desired answer to any specific question. This does not imply a simple one-size-fits-all process, rather that the logic for modifying an analysis for specific research questions should be explicated and applied systematically. A im: Avoids P hacking and allows for the evaluation of the operating characteristics of this process (Principle 6).

5 LEGEND will generate evidence using best practices.

LEGEND answers each question using current best practices, including advanced methods to address confounding, such as propensity scores. Specifically, we will not employ suboptimal methods (in terms of bias) to achieve better computational efficiency. Aim: Minimize bias.

6 LEGEND will include empirical evaluation through the use of control questions.

Every LEGEND study includes control questions. Control questions are questions where the answer is known. These allow for measuring the operating characteristics of our systematic process, including residual bias. We subsequently account for this observed residual bias in our P values, effect estimates, and confidence intervals using empirical calibration[7,8]

Aim: Enhances transparency on the uncertainty due to residual bias.

- 7 LEGEND will generate evidence using open-source software that is freely available to all. The analysis software is open to review and evaluation, and is available for replicating analyses down to the smallest detail. Aim: Enhances transarence and allows replication.
- 8 LEGEND will not be used to evaluate new methods.

Even though the same infrastructure used in LEGEND may also be used to evaluate new causal inference methods, generating clinical evidence should not be performed at the same time as method evaluation. This is a corollary of Principle 5, since a new method that still requires evaluation cannot already be best practice. Also, generating evidence with unproven methods can hamper the interpretability of the clinical results. Note that LEGEND does evaluate how well the methods it uses perform in the specific context of the questions and data used in a LEGEND study (Principle 6).

Aim: Avoids bias and improves interpretability

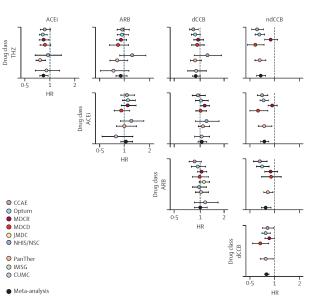
9 LEGEND will generate evidence across a network of multiple databases.

Multiple heterogeneous databases (different data capture processes, health-care systems, and populations) will be used to generate the evidence to allow an assessment of the replicability of findings across sites.

Aim: Enhances generalizability and uncovers potential between-site heterogeneity.

10 LEGEND will maintain data confidentiality; patient-level data will not be shared between sites in the network. Not sharing data will ensure patient privacy, and comply with local data governance rules. Aim: Privacy.

Note: LEGEND: Large-scale Evidence Generation and Evaluation across a Network of Databases.





2. Objective Diagnostics

Engineering open science systems that build trust into the real-world evidence generation and dissemination process

Required phenotypes Analysis specifications Deskipe kerkelde	dardized to common data model
stop definitions diagnostics	Analysis reliability evaluation
System characteristics: Standardized procedures with defined inputs and outputs Analysis packages implementing scientific best practices consistently applied across all data partners, generating consistent	choice Fail
output for network synthesis Reproducible outputs generated by open-source analysis libraries developed and validated with verifiable unit-test coverage Pre-specified and objective decision thresholds for go/no go criteria Measurable operating characteristics of system performance	unblinded results interface exploration

Statistical power Rule: Minimum Detectable Relative Risk (MDRR) < 10

Comparability Rule: Equipoise > 0.5

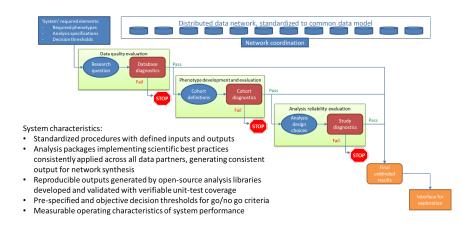
Covariate Balance Rule: Max standardized difference of mean (SDM) < 0.1

Generalizability Rule: Max SDM between analytic cohort and target cohort < 0.25

Residual systematic error Rule: Expected Absolute Systematic Error (EASE) < 0.25



Statistical power rule



Hazard Ratio (95% Cl) 4.06 (0.39 - 42.60) 1.05 (0.86 - 1.28) 0.1 0.25 0.5 1 2 4 6 8 10 Hazard Ratio

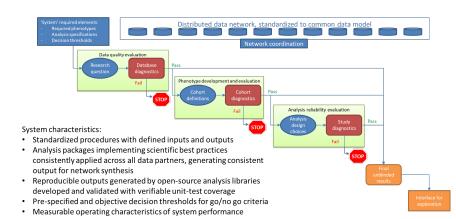
Rule: Minimum Detectable Relative Risk (MDRR) < 10

Reasoning: Even low-power estimate (wide CI) could be helpful, but we want to avoid misinterpreting grossly underpowered studies

Minimum Detectable Risk Ratio (MDRR) is a term used to describe the smallest relative risk that a study with a given power is capable of detecting.



Comparability Rule (Equipoise)



Equipoise = 83% 0 0.5 1 0 0.5 Preference score Equipoise = 28% 0 0.5 Preference score

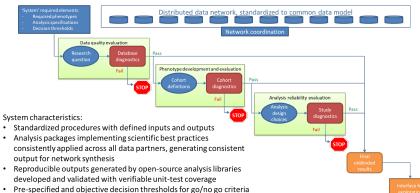
Rule: Equipoise > 0.5 (Equipoise is percent of population that has 0.3 < preference score < 0.7)

Reasoning: If equipoise is low, the populations are too incomparable, and we probably shouldn't trust our ability to make them comparable.

Preference = probability of patient choosing target vs. comparator treatment, given baseline features

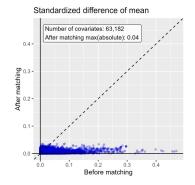








Exposure $\xrightarrow{\text{Effect of interest}}_{\text{RR=???}}$ Outcome



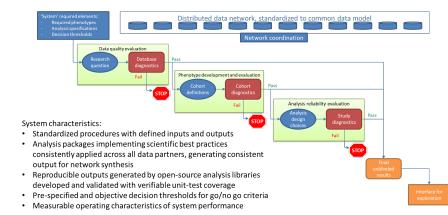
Rule: Max standardized difference of mean (SDM) < 0.1 (no covariate may have a SDM >= 0.1 after PS adjustment)

Reasoning: If covariates are unbalanced there may be confounding.

Confounding variables associated with both exposure and outcome can bias effect estimates if not properly addressed



Generalizability rule



Strategies employed to reduce confounding (e.g. PS matching) can shift the composition of the analytic cohort away from the original target

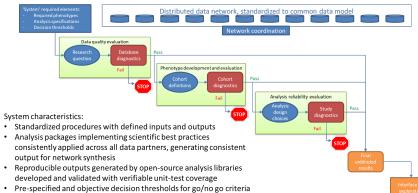
Rule: Max SDM between analytic cohort and target cohort < 0.25

(target cohort: the cohort we started with (those exposed)) (analytic cohort: the cohort after all adjustments)

Reasoning: Estimate may not generalize to our target population if differences are too great.

Generalizability is the extent to which a study result can be applied to a target population of interest

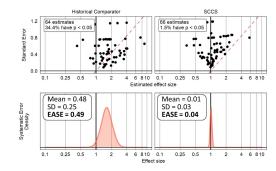




• Measurable operating characteristics of system performance

Bias – expected value of systematic error – can be estimated using negative control experiments in which estimates can be compared with known truth

Residual systematic error rule (EASE statistic)



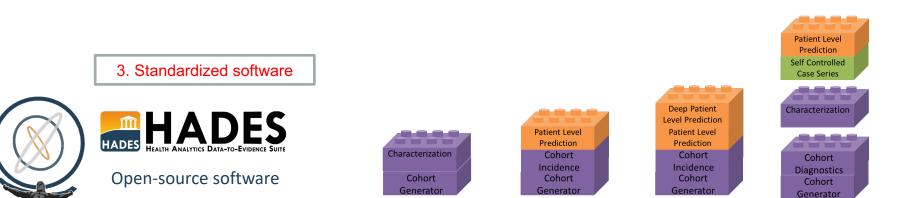
Rule: Expected Absolute Systematic Error (EASE) < 0.25 (EASE is the expected abs(log(estimated RR) – log(true RR)), based on negative control estimates)

Reasoning: Even though we can and should empirically calibrate to account for residual error, readers may not trust results if calibration shifts the estimates too much.

Residual systematic error can still exist due to model misspecification inherent to analysis or data

TRUST 3: Standardized software





Cohort Generator: R package for generating cohorts using data in the CDM Cohort Diagnostics: Evaluation of phenotype algorithms for OMOP CDM Cohort Incidence: Performs incidence calculations on a CDM Characterization: Performs characterization on target and comparator cohort Cohort Method: performs new-user cohort studies in the OMOP CDM Self Controlled Method: Performs Self-Controlled Case Series (SCCS) analyses in the OMOP CDM Patient-Level Prediction: Performs patient level prediction in the OMOP CDM Evidence Synthesis: R package for combining evidence from multiple sources

TRUST 3: Standardized software



What is Health Analytics Data to Evidence Suite (HADES)

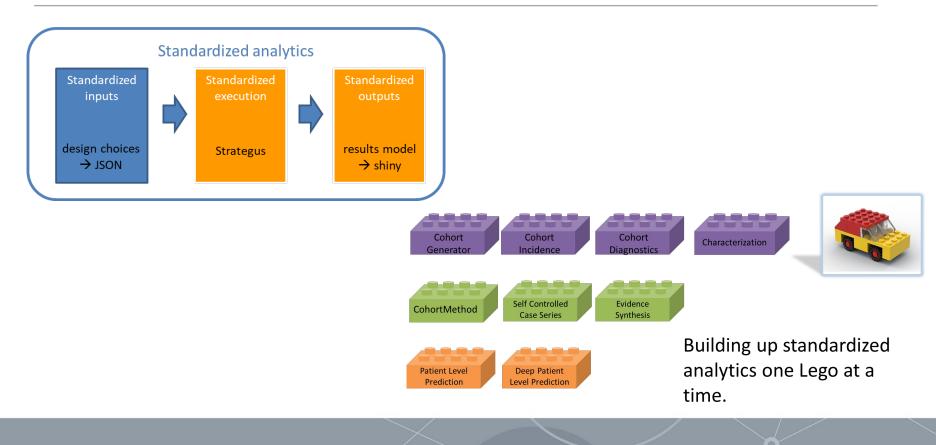
- Open-source R packages for execution on OMOP CDM
- Principled software, grounded in methods research
- Documented, maintained, tested, empirically validated software
- Facilitates multi-question
- Facilitates large-scale analytics (big data)



- Distributed data network support: Enables federated analyses
- Platform independent: Compatible with diverse technical infrastructures

TRUST 3: Standardized software





Conclusion: TRUST \rightarrow TCIO-TAR



Three ideas

2. Objective Diagnostics

Phenotype development and evaluation

Fail

Engineering open science systems that build trust into the real-world evidence generation and dissemination process

Distributed data network, standardized to common data model

Network coordination

Analysis reliability evaluation

3. Standardized software



Open-source software



System characteristics:

Standardized procedures with defined inputs and outputs

Data quality evaluation

- Analysis packages implementing scientific best practices
- consistently applied across all data partners, generating consistent output for network synthesis
- Reproducible outputs generated by open-source analysis libraries developed and validated with verifiable unit-test coverage
- Pre-specified and objective decision thresholds for go/no go criteria
- Measurable operating characteristics of system performance



Principles of Large-scale Evidence Generation and on many outcomes). Aim: Avoids publication bias, Evaluation across a Network of Databases (LEGEND) achieves commehensiveness of results, and allows for Martijn J. Schuemie (3^{1,2}, Patrick B. Ryan^{1,3}, Nicole Pratt⁴, RuiJun Chen (3^{1,5}, Seng Chan You⁶, Harlan M. Krumholz⁷, David Madigan⁶, George Hripcsak^{2,6}, and an evaluation of the overall coherence and consistency Marc A. Suchard^{2,1}

2. Dissemination of the evidence will not depend on the estimated effects. All generated evidence is disseminated at once. All mit works within this and enhances transparent

of the generated evidence.

3. LEGEND will generate evidence using a prespecified analysis design. All analyses, including the research questions that will be answered, will be decided prior to analysis execution. Aim: Avoids P hacking

1. LEGEND Principles

METHODS RESEARCH

4. LEGEND will generate evidence by consistently applying a systematic process across all research questions. This principle precludes modification of analyses to obtain a desired answer to any specific question. This does not imply a simple one-size-fits-all process, rather that the logic for modifying an analysis for specific research questions should be explicated and applied systematically. Aim: Avoids P hacking and allows for the evaluation of the operating characteristics of this process (Principle 6)

5. LEGEND will generate evidence using best practices. LEGEND answers each question using current best practices, including advanced methods to address confounding, such as propensity scores. Specifically, we will not employ suboptimal methods (in terms of bias) to achieve better computational efficiency. Aim: Minimizes bias

6. LEGEND will include empirical evaluation through the use of control questions. Every LEGEND study includes control questions. Control questions are questions where the answer is known. These allow for measuring the operating characteristics of our systematic process, including residual bias. We subsequently account for this observed residual bias in our P values, effect estimates, and confidence intervals using empirical calibration. [7.8] Aim: Enhances transparency on the uncertainty due to residual bias

7. LEGEND will generate evidence using open-source software that is freely available to all. The analysis software is com to review and evaluation and is available for replicating analyses down to the smallest detail Aim; Enhances transparency and allows replication

8. LEGEND will not be used to evaluate new methods. Even though the same intrastructure used in LEGEND may also be used to evaluate new causal inference methods, generating clinical evidence should not be performed at the same time as method evaluation. This is a corollary of Principle 5, since a new method that still requires evaluation cannot already be best practice. Also, generating evidence with unproven methods can harmore the interpretability of the clinical results. Note that I EGEND does evaluate how well the methods it uses perform in the specific context of the questions and data used in a LEGEND study (Principle 6). Aim: Avoids bias and improves interpretability.

9. LEGEND will generate evidence across a network of multiple databases. Multiple heterogeneous databases (different data capture processes, health care systems, and populations) will be used to generate the evidence to allow an assessment of the replicability of findings across sites. Alm: Enhances generalizability and uncovers potential between-site heterogeneity.

10. LEGEND will maintain data confidentiality; patient-level data will not be shared between sites in the network Not sharing data will ensure patient privacy, and comply with local data governance rules. Aim: Privacy,

#JoinTheJourney	47	OHDSI.org
-----------------	----	-----------





Thank you!

Email me at: rao@ohdsi.org

Agenda





Reproducibility and Trust

Ross D. Williams, PhD Erasmus University Medical Center Darwin EU[®] Analytics Team Lead r.williams@erasmusmc.nl





This presentation represents the views of the DARWIN EU® Coordination Centre only and cannot be interpreted as reflecting those of the European Medicines Agency or the European Medicines Regulatory Network.

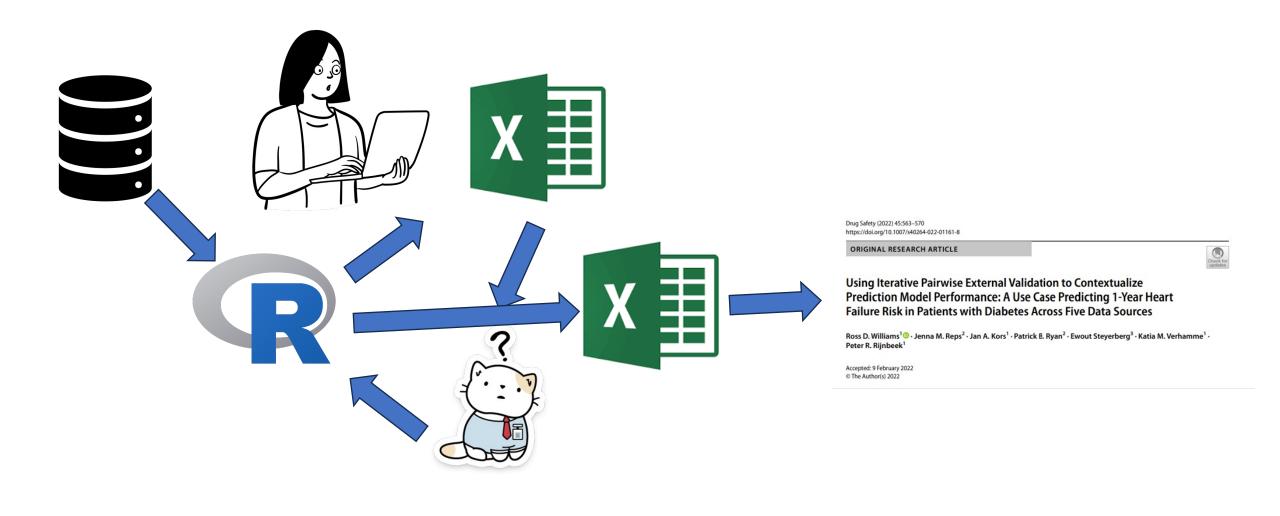


What do we mean by reproducibility?

Desired						
attribute	Question	Researcher	Data	Analysis		Result
Repeatable	Identical	Identical	Identical	Identical	=	Identical
Reproducible	Identical	Different	Identical	Identical	=	Identical
Replicable	Identical	Same or different	Similar	Identical	=	Similar
Generalizable	Identical	Same or different	Different	Identical	=	Similar
Robust	Identical	Same or different	Same or different	Different	=	Similar
Calibrated	Similar	Identical	Identical	Identical	=	Statistically
	(controls)					consistent



What does a traditional epi study look like?

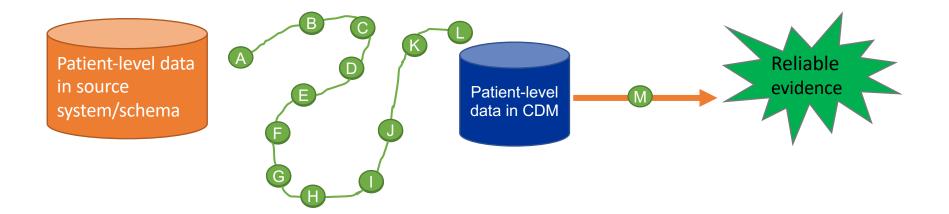






Generating Reliable Evidence using the OMOP Common Data Model

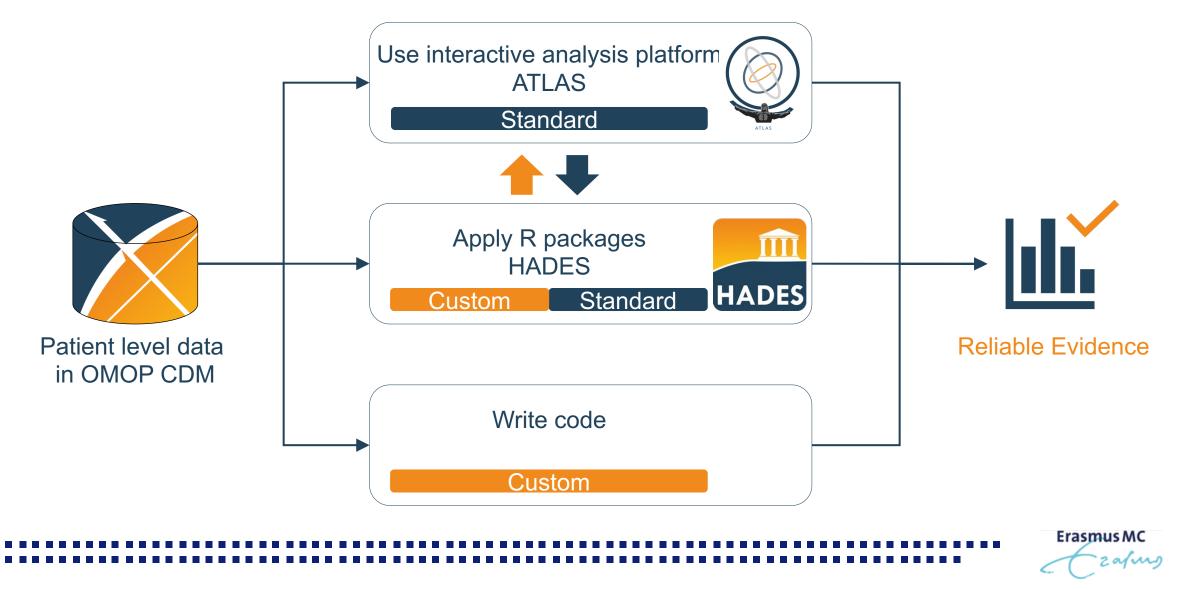
We need to make studies repeatable, reproducible, replicable, generalisable, and robust



A Common Data Model enables standardised analytics to generate reliable evidence.



OHDSI roads to reliable evidence



Diagnostics

- By standardizing the elements of studies (Characterisation, Estimation, Prediction)
 - We can standardize diagnostics





Darwin EU® perspective

- Standardise the questions, this allows for standardised software
- For the standard questions have standard software
 - Test this software
 - When errors are discovered, create a test, fix error
- This produces better software, more reliable answers and the research is reproducible
- Use renv to increase the likelihood of reproducing estimates at a later point



Software Review Process





Clarity and openness

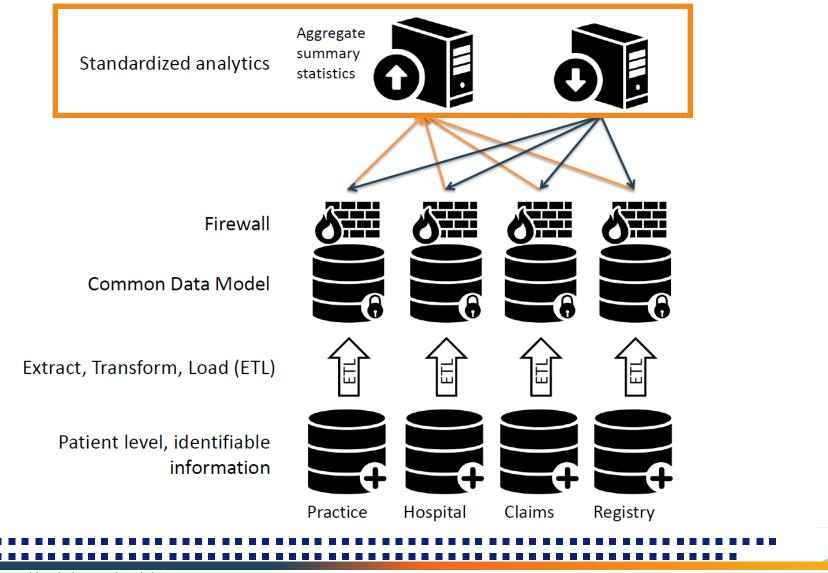
We have to respect patient privacy, and we must be open with analysis

Protocols, standard software, clear decision making





Federated analyses in OHDSI network



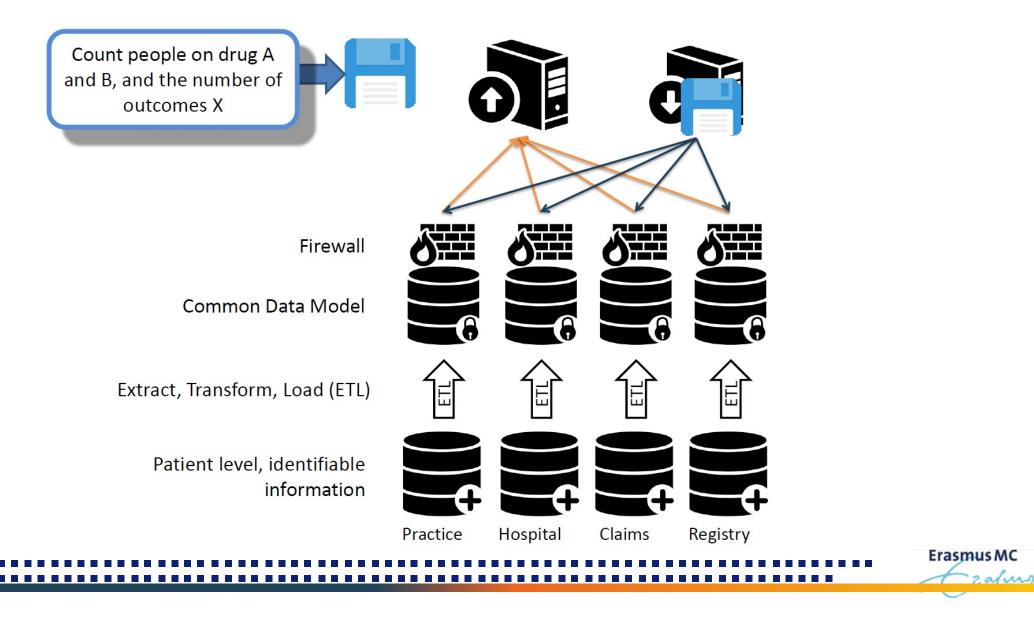
Schuemie MJ. How to extract transform and load observational data.

https://www.ohdsi.org/wp-content/uploads/2014/07/Beijing2015.pdf? ga=2.178811554.749634320.1678273784-1300990784.1664885317 Last accessed 09-MAR-23

Erasmus MC

zalus

Federated analyses in OHDSI network



Why standardisation makes research more trustworthy

Standardised pipelines can be incrementally improved over time

Flexibility can be improved through user interaction and development cycles

Moving to a standard design approach

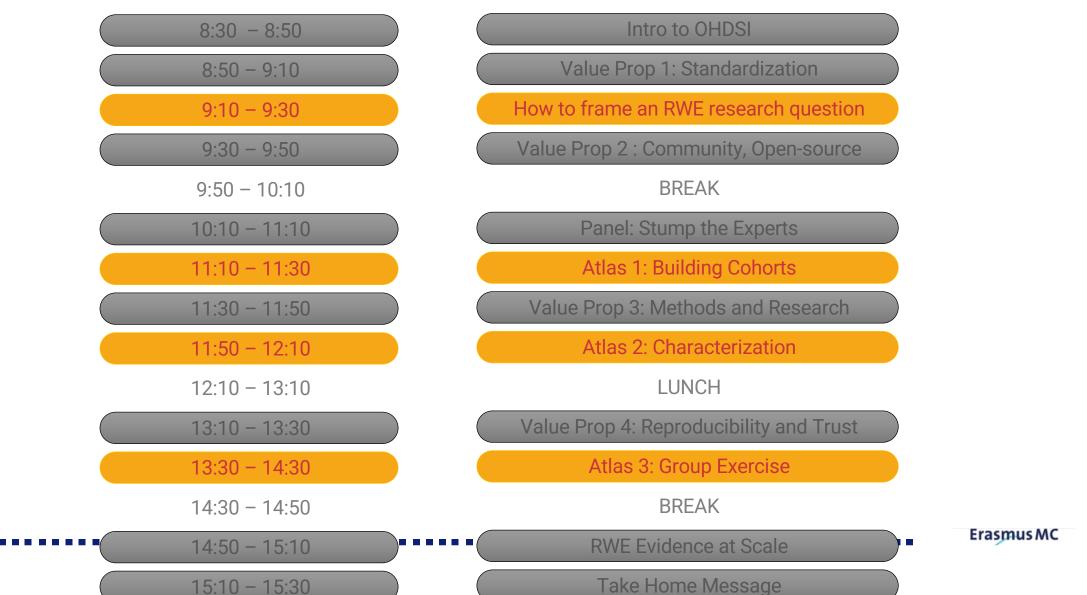




Thanks for your attention r.williams@erasmusmc.nl









Evidence generation at scale

11-Nov-2023





I disclose the following relevant relationship with commercial interests:

• VP, Head of Data Science at Odysseus Data Services



Current pace of evidence generation in healthcare



All health outcomes of interest

		BI Ca oordi	Co Ea En I	Eye Gastro disor testina			Infectio	ons and		soning Investig dura ions		is Musculos				Nerve	ous system	Pregnancy. puerperium an	Psychiatr	ic Renal an urinary d	d Reprod	ductive	Respiratory, thoracic and me.	Skin and	Social Sur	gical med Vascular disorders
		M				All Mid B				Oc Pr S H N	Bo Ir O	B In Joi Mu	O VI I	H Ly M N	PVIC	r In M N	NN Pe S		r C M N S	S Inj R U	Jr Br M P	R Re Ut	In PIR R SIL	CE LISK	Sk Vi Le H 1	Th'CGN Sk Va Vas Va Va
		teres a		n n n n n N n n ngah				1 4 4		n antar a cara a ca Cara a cara a	4.55	1	1.6				1.1.1	1.1	1.1.1.1.1		1.1.1.1	1.1	 I definition (equipal) 	a mante	979 F L 1	American American
	OTHER INT	s de se	1.0	1.1.1.1	14.4			1.11	1.1	1.1	dia dia	a - 6			1 de 1	i n	and a start	and the second	diment of	0.11	1.00		11 a	a de la		with the second second
	AGID PREP	1000		10.01	1.0	na de	1.1	1.1.1	1.00	1.1	one's	1.1.2	16	1 - B		diana.	nine i	19.00 C	100.000	a de	32.00	1.16	v v v	10000	11	and some second
	PROPULSI	1.1		10 C 11 M							all a		40.0	1.1			P (1 = 4	1	100	1.11	1.1	. T.	1.18	1.01		
	OTHER BL	inge-	- N	1.1.1.1	19 B	ê na se			1.0		- 29 1 -27	1990 - 1940 - 1940 - 1940 - 1940 - 1940 - 1940 - 1940 - 1940 - 1940 - 1940 - 1940 - 1940 - 1940 - 1940 - 1940 -	1.1	1912		2010		19 March 19	4.5.5	计分析		1.1	6 A 9 A 1	dikter -	6.57	이 말 가 있어요? 가지
	OTHER MIN	1.1		M			1	1.1		() (i)	40.0	1.1.1.1				1.1	1.11	1.1	1.1	$(x_{i}) \in \mathbb{Z}$	1.5		1.1.1.1.1.1.1	1.1		and the state of the
	ENZYMES-	66.	1.1	Carlei M	12.4	N.,				1.1	Same	end i i	10.1		1.1	Sec.	tapan d		- 61 T B	1 A.	1.1		1.2.4.4.4.4	A REAL PROPERTY.	en l	가지 지난 바이
ĩ	NTERMEDI	出版。	1.1	1. 1. 1. 44	14	and a state		18 A.		i i i	A.	a da	1	1.0	- 1 A.	dise.	alard.	1.1.1	all a s	- 2 K	1.6	1.1	Back to a	1486	1	i i a tasti da
	SECOND-G	110		da ti				1011			- 1040 - 2010 - 2010	la de	1. s		1	12.1	201	11 A.A.A.A.A.A.A.A.A.A.A.A.A.A.A.A.A.A.A		- 91.	3.4.1		i di second	1 Editor -	.a. –	그는 말을 하는 것 같아.
	HYDRAZIDES -	122	1999 - 1999 1999 - 1999 - 1999 - 1999 - 1999 - 1999 - 1999 - 1999 - 1999 - 1999 - 1999 - 1999 - 1999 - 1999 - 1999 - 1999	a she a shi a	- 1 B					1.1	- 111 (12) - 212 (12)	655 - 1935 6 106 - 1945	21 - E	2.1		$E_{\rm s}^{\rm ext}$	22221		藏計	1.1.1	100		in a di data i Na dia kaominina	i o rite na 11 Tata a la casa	F 2 F	and the second second
1	ROTA VIRU	110.0		N = 1.88 N = 1.44		The second second	16.10		1.1	1.8	a da a com		Нτ.,	and in		1979.		1.11		나라 같다.			an a statuti da el	and the second second second	5 A.	 A state of the sta
	OTHER AL ANTI-ESTR	- 1995		1 - Al-	e al de	1940 C.	1.11	61.11	1.11	£166 - 1		dates in the	1.	10^{-1}				1994 - A	1822	0.94	199 A.S.		에 비행기록 가지	ald pages.	ter -	and we could be of a
ť.	BENZIMIDA	100	Contract of the	ti di	n niniir	station of the		fil serve	de la composición de	2.44	- Hereit		12.2		11				15674	1.111			il detter	ang ang pa	istr.	saniti ng aga siya siya biran
	BIGUANIDES - RON IN CO	- 88.5	111.0	78 11	19.6	111		1.414	1.1	1.00	- 89	1. A 199 A		98 a 1	1.1	1105.	1.4.6.6.7	- 60 (P	1492.1			. A.	80.44 s	6.63.67	5 y	NAB KAN
1	VITAMIN K-	14.1			1.00					1.1		1.11				P = 2	1100		10	1.11	1.0	1				2011 C C C C C C C C C C C C C C C C C C
	MAGNESIUM-	123	e de la composición d La composición de la c	111 日前	1 12 12 12 12 12 12 12 12 12 12 12 12 12	Aller St.		18.	1.1	1.1	頭筋	en en presente de la composición de la En este de la composición de la composic	1.1	1.00		10.00	-195 A (Record and the second se	12.11	and the second sec	1.1	1	n andre and An an an an	e isie al Salutz		nandigi magi sa mana dali kadi ng Kin Takang na mang pang takang ng King ng Kang
	OTHER AN	-642		- A - 19.	6	dia 🖓		1.11		1	200	ar 1914	14.1		1.1	1.44	1414	1.00	allen i	3 N.F.	S		Ne Miller	140	14	
	ETA BLOC	. da P	1.1.1.1	inter a dag	11.0	÷		1.11	1.00	1.	- sili (s.	11.11		1.0	[1.25	- 28月1日月	6.2012	- this does	an tik	电正规	3	P 6 19 1 1 1 1	that the second	#11	网络白垩属白垩合合
	ANTIARRH	- 29 (2	5 A. A.	20 M	164	40 H.	÷	à an		1.1	- Wester	5 B.	81		1 de 1	1. 11	Ricci	10 C 10	- Made C		생활하는		10 - 20 - 5 11 - 20 - 5 - 1	1990 - S	<u>9</u>	해당 수영되는
-{-	HIAZIDES	- 948 2	1.1	y i pag	主義	and the second			1.1	- i	- 499-144		31		1.1	for t	- 660 A - 8	19 B	4.14	计时间	-94 T - 1		distriction of the	1. Contraction	5 y	
	2-AMINO-1 BIOFLAVO	- 327		3 - BP	i ya w	1.1		1.1		1.1	1993	1.1.1	1.0	100	· · ·	47.1		11.1	- <u>1</u>	1.112	100	1.1		d refer o		[1] J. S. K. K. M.
	OTHER AN	- 66 s	1.1	动行机	- X B		12.00		1.00	1.1			11 A.	100	6 - A.	11.1	- ANT - S	14.1	1997	1.11	11000		N 1 1 1 1 1	The second se	20	 Bernsteiner auf der Aussellungen der Aussellung Aussellungen der Aussellungen der Aussellungen
	NUCLEOSI	្រំដែរច	- 1 C - 1	1.14	1.1			- C	1.1	1.1	- 419	49		1	1.1	frank.	err si se	dia	나랍하는	1 A.	- 40 - C		如果我们的	「翻訳」、	21	والمشتر والمساور
	ANTIPROG	122		1 - 11	- 12 B	4 C 1		1.00		1.12	- 36 ° 1		10	9 - C.	÷	42.27	11.00	100 A. C.	an Car	1.11	NI 47	11.24	1997 - N. S.	Sec. 2	n (* 1997)	이야한 이 문화가 된다.
	ELECTIVE	Mas	18 J. 1	1.832	18.4	14 A A A A		V.		1.1	1810	a aga		syt og t		10.15	학생님	- 34 A.	1951 4	(1 + y)	تىرە خەرى	en fet	te Maria de	-1111 (C	ed in	adu gʻugʻt Kigʻo
	COXIBS -	ila! :	1.1	10.00	1 56	a tr			1.1	14	- Barr	67 a. G	÷ .	6.1.5		4134	- 3 C + 1	de la composición de la compos	11991	生放转	생활이 있는	1.3	Sec. Beech	田田一	$C_{i} = 0$	rada Merci Mirishi
	ROPIONIC	108.7	1.16	11日 福祉	104	14 B		1.11.1	1.11	1 B - 1	a silê A s	en far	94 L -	12.00	1.1	dia B	194701-1	al la co	- 640 - E	a di te	1.00	10.00	推測者主要の	場 監督部 いっ	949 - C	(2) 日本の間によい。
	OTHER CE BENZOMO	diffe in		A State	n li	4.8	6. A. A.	6.000		1.1			6			APRIL 1	1884 - E	dia anti-	na an taon an t	一 山田			14414	dition at	¥1.	and the approximation
4	ELECTIVE	- 33.7	. N. M	문 집	清靜	e de la composition d		200	14.9	1.	관람은	11 D B.	G1 .		inte i	là di	103 - en el 1 1112 - en el 1	10.11	all the second	8 H	Sec. As a	- L -	12월 11일	2.666	5 - 875 C	1111日日111日
	OTHER AN	- 44,5	11.1.1.1.10	- 1 M	1 C 10	10.0		M	1111	19 - Marcola	Sector Sector	an sag	1.1	- 2	1.1	16.41	潮線出	이번 모두 문		化动物	2.1	- R4	이 많은 것이.	a (1997)	1973 - A	그 방법 요즘은 모두는
	PROLACTI	- 11 li 1	1.11	وليهدد والدفر	- t 4:	a na she	1.1		1.0	1.11	and the second		111	222	9 H -	de de	1921 - 1	- 11 (L. 11)	말한 날리	il) e y	2112.2	1912-1	n a tha an s		1960 (N	and second second
	OTHER NE		1.1	N	1.5		1.1		111		42.1		<u>.</u>			17.31	di Anti	12 March 19	HID CLUB	an sh	10.00	1.1	80 3 0 6 -	1.26%	8	A Charles and a charles
	ALDEHYDE	- 1 18 6	1 a a 4	1.1	10.50	3 (p)		1.1			100.07	ene tál	i. –			an 1997 -	- 時間 ほうよう	Statistics -	雕場	14 - C.L.	164 A.Y	1.4.4	1.0.00110.0		900 - C	湖口是建筑时代。
	CARBAMAT	- 18 -	- J.	21日 静	马根			11 a A			Siles			88 ⁽¹ 1)	11	招供	應用	생활	贈出	19 T.	1811 E.		新聞 (1997年)	A DECKER OF	11 1	그래픽 실망로 가는
	MELATONI PHENOTHI	1986	the file	101 - 51	지하	nde l'		1.1		1.1	THE ST.	S to po	Sec. 1	1.1	· .	11.41	100-01-0	gr its	100.00	g2 - 1 -	der ta e	1.1	ha da a la pa	新聞的 人人	Market I.	्योंको के किस्तु न
	SUBSTITUT			$A_{ij} \in A_{ij}$	n de de			1.00		1.1	14-15		÷			45 B	1966	1.1	- 바람 - 문제	9 C A	1.1		1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 -	$1 \ \{0, \dots, n\}$	14	and the second
	OPIUM ALK SYMPATHO	1.54		1 1 44	16.8	Sacre.		1.1	1.1	1.1	1811.	91 - A.	1		6 - C	1.52	111	1.1.1	Here La L	11.511	100	1.1	间端的	14-04-04-04-04 		and the second
	OTHER AN	1281		11.15	対象				. e	8 (j. 1	出出。	an di	15.5			10^{-12}	1994	1.10	网络牙子	1.11	19.20		1. 18 de l	114	12	and the second second
	SYMPATHO	100	18 A.	Samal		ang in		9.6		1.16	and the second	a da	a di			fact.	描述目	t agus fair	- 1993 - C	98 (H)	4.42	160	11811	A DOUBLE AN	Geo.	- 推动 - 网络花椒
	GLUCOCO	-496	1 N - 🖞	구성 관련	日連	4 H. A.	1 - 1 - 1	politica.	11.1	-14 C	े समें स्ट	filend a	a a a c	eter de	14.5	44	田柏田	18 A.	्राईका व	法出职	deres.	- ÷+	3 ∩ A ⊕ E ⊕	1222	p i e	油酸钙 网络白垩
	ANTIDOTES -	- 5° 1	12.14	计分析	- h 4	Bi (A.C.	1.1	1.11		. C	- <u>1</u> 21 - 1	100 J.Y.	111			1978	25.001	1111	100.4	22 H	11	1	9 M H H H H H	김 관리가?	2	그 없는 것 같은 것이 같다.
-	VATERSOL	- 240	1.1	t deb	u b	1.22	1.00	1.11		1.00	POP 4 11	M (187	211	1.1	1.1	1.11	18.0 B	1	tinin t	n n fe	1.1		2015-00	110	9 F.	고 한 곳 11 (12) 환자

All drugs

Current evidence base for hypertension



Head-to-head antihypertensive drug comparisons



- Driven primarily by one clinical study ALLHAT- only 3 individual drugs
- Focus: mostly on efficacy

Can we provide

- 1. reliable concordant w/ RCTs
- **2. rich** across "all" comparators
- **3. relevant** inform practice evidence?







HEALTH CARE REFORM

Comparative Effectiveness of 2 β -Blockers in Hypertensive Patients

Emily D. Parker, MPH, PhD; Karen L. Margolis, MD, MPH; Nicole K. Trower, BS; David. J. Magid, MD, MPH; Heather M. Tavel, BS; Susan M. Shetterly, MS; P. Michael Ho, MD, PhD; Bix E. Swain, MS; Patrick J. O'Connor, MD, MPH



Two targets: atenolol and metoprolol Three outcomes:

- Acute myocardial infarction
- Stroke
- Heart failure

ARCH INTERN MED/ VOL 172 (NO. 18)

Single ingredient comparisons

	Theoretical	Observed (n>2,500)
Single ingredients	58	39
Single ingredient comparisons	58 * 57 = 3,306	1,296





+ Single drug classes comparisons

	Theoretical	Observed (n>2,500)
Single ingredients	58	39
Single ingredient comparisons	58 * 57 = 3,306	1,296
Single drug classes	15	13
Single class comparisons	15 * 14 = 210	156





+ single vs dual ingredient comparisons

	Theoretical	Observed (n>2,500)
Single ingredients	58	39
Single ingredient comparisons	58 * 57 = 3,306	1,296
Single drug classes	15	13
Single class comparisons	15 * 14 = 210	156
Dual ingredients	58 * 57 / 2 = 1,653	58
Single vs duo drug comparisons	58 * 1,653 = 95,874	3,810





+ dual classes comparisons

	Theoretical	Observed (n>2,500)
Single ingredients	58	39
Single ingredient comparisons	58 * 57 = 3,306	1,296
Single drug classes	15	13
Single class comparisons	15 * 14 = 210	156
Dual ingredients	58 * 57 / 2 = 1,653	58
Single vs duo drug comparisons	58 * 1,653 = 95,874	3,810
Dual classes	15 * 14 / 2 = 105	32





+ single vs dual class comparisons

	Theoretical	Observed (n>2,500)
Single ingredients	58	39
Single ingredient comparisons	58 * 57 = 3,306	1,296
Single drug classes	15	13
Single class comparisons	15 * 14 = 210	156
Dual ingredients	58 * 57 / 2 = 1,653	58
Single vs duo drug comparisons	58 * 1,653 = 95,874	3,810
Dual classes	15 * 14 / 2 = 105	32
Single vs duo class comparisons	15 * 105 = 1,575	832





+ dual vs duo drugs comparisons

	Theoretical	Observed (n>2,500)
Single ingredients	58	39
Single ingredient comparisons	58 * 57 = 3,306	1,296
Single drug classes	15	13
Single class comparisons	15 * 14 = 210	156
Dual ingredients	58 * 57 / 2 = 1,653	58
Single vs duo drug comparisons	58 * 1,653 = 95,874	3,810
Dual classes	15 * 14 / 2 = 105	32
Single vs duo class comparisons	15 * 105 = 1,575	832
Duo vs duo drug comparisons	1,653 * 1,652 = 2,730,756	2,784





+ dual vs duo class comparisons

	Theoretical	Observed (n>2,500)
Single ingredients	58	39
Single ingredient comparisons	58 * 57 = 3,306	1,296
Single drug classes	15	13
Single class comparisons	15 * 14 = 210	156
Dual ingredients	58 * 57 / 2 = 1,653	58
Single vs duo drug comparisons	58 * 1,653 = 95,874	3,810
Dual classes	15 * 14 / 2 = 105	32
Single vs duo class comparisons	15 * 105 = 1,575	832
Duo vs duo drug comparisons	1,653 * 1,652 = 2,730,756	2,784
Duo vs duo class comparisons	105 * 104 = 10,920	992





+ dual vs duo class comparisons

	Theoretical	Observed (n>2,500)
Single ingredients	58	39
Single ingredient comparisons	58 * 57 = 3,306	1,296
Single drug classes	15	13
Single class comparisons	15 * 14 = 210	156
Dual ingredients	58 * 57 / 2 = 1,653	58
Single vs duo drug comparisons	58 * 1,653 = 95,874	3,810
Dual classes	15 * 14 / 2 = 105	32
Single vs duo class comparisons	15 * 105 = 1,575	832
Duo vs duo drug comparisons	1,653 * 1,652 = 2,730,756	2,784
Duo vs duo class comparisons	105 * 104 = 10,920	992
Total comparisons	2,843,250	10,278





+ expert curated outcomes

	Theoretical	Observed (n>2,500)
Single ingredients	58	39
Single ingredient comparisons	58 * 57 = 3,306	1,296
Single drug classes	15	13
Single class comparisons	15 * 14 = 210	156
Dual ingredients	58 * 57 / 2 = 1,653	58
Single vs duo drug comparisons	58 * 1,653 = 95,874	3,810
Dual classes	15 * 14 / 2 = 105	32
Single vs duo class comparisons	15 * 105 = 1,575	832
Duo vs duo drug comparisons	1,653 * 1,652 = 2,730,756	2,784
Duo vs duo class comparisons	105 * 104 = 10,920	992
Total comparisons	2,843,250	10,278
Outcomes of interest	58	58
Target-comparator-outcomes	2,843,250 * 58 = 164,908,500	587,020





Creating an evidence base for hypertension + Diagnostics

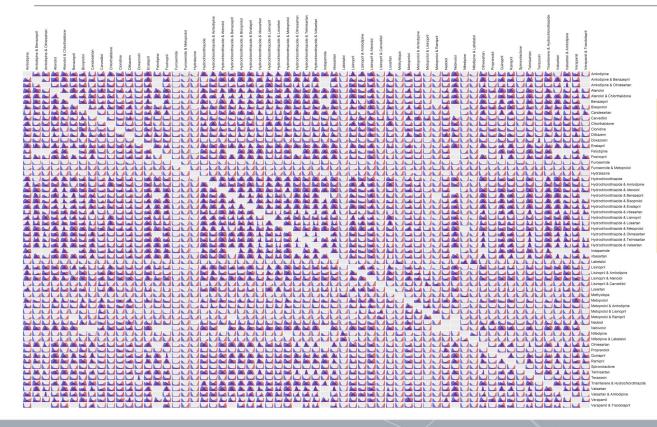
	Theoretical	Observed (n>2,500)
Single ingredients	58	39
Single ingredient comparisons	58 * 57 = 3,306	1,296
Single drug classes	15	13
Single class comparisons	15 * 14 = 210	156
Dual ingredients	58 * 57 / 2 = 1,653	58
Single vs duo drug comparisons	58 * 1,653 = 95,874	3,810
Dual classes	15 * 14 / 2 = 105	32
Single vs duo class comparisons	15 * 105 = 1,575	832
Duo vs duo drug comparisons	1,653 * 1,652 = 2,730,756	2,784
Duo vs duo class comparisons	105 * 104 = 10,920	992
Total comparisons	2,843,250	10,278
Outcomes of interest	58	58
Target-comparator-outcomes	2,843,250 * 58 = 164,908,500	587,020
Diagnostics	164,908,500	587,020





Best-practices: systematic large-scale PS



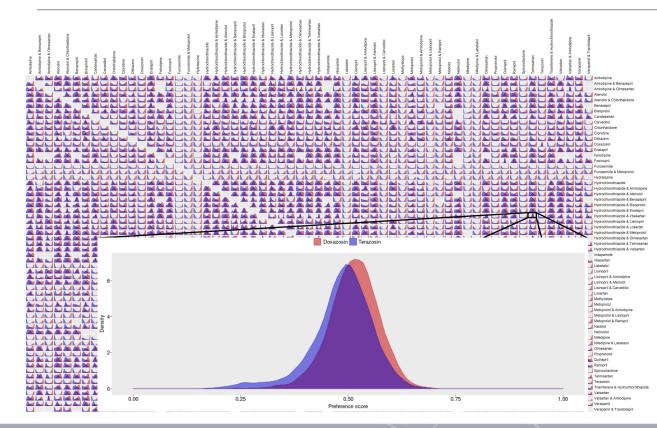


- >8,000 (regularized) baseline patient characteristics (all dx, rx, tx)
- Address observed (and some unobserved – BP control) confounding

Tian et al, 2019, IJE

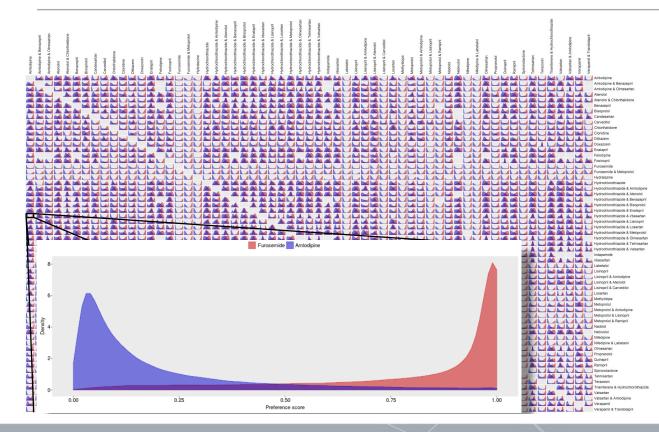
Not all comparisons are valid





Not all comparisons are valid





+ negative controls

Theoretical	Observed (n>2,500)
58	39
58 * 57 = 3,306	1,296
15	13
15 * 14 = 210	156
58 * 57 / 2 = 1,653	58
58 * 1,653 = 95,874	3,810
15 * 14 / 2 = 105	32
15 * 105 = 1,575	832
1,653 * 1,652 = 2,730,756	2,784
105 * 104 = 10,920	992
2,843,250	10,278
58	58
2,843,250 * 58 = 164,908,500	587,020
76	76
2,843,250* 76=216,087,000	769,476
	58 58*57=3,306 15 15*14=210 58*57/2=1,653 58*57/2=1,653 15*105=95,874 15*105=1,575 1,653*1,652=2,730,756 105*104=10,920 2,843,250 58 2,843,250*58=164,908,500 76





+ positive controls

	Theoretical	Observed (n>2,500)
Single ingredients	58	39
Single ingredient comparisons	58 * 57 = 3,306	1,296
Single drug classes	15	13
Single class comparisons	15 * 14 = 210	156
Dual ingredients	58 * 57 / 2 = 1,653	58
Single vs duo drug comparisons	58 * 1,653 = 95,874	3,810
Dual classes	15 * 14 / 2 = 105	32
Single vs duo class comparisons	15 * 105 = 1,575	832
Duo vs duo drug comparisons	1,653 * 1,652 = 2,730,756	2,784
Duo vs duo class comparisons	105 * 104 = 10,920	992
Total comparisons	2,843,250	10,278
Outcomes of interest	58	58
Target-comparator-outcomes	2,843,250 * 58 = 164,908,500	587,020
Negative control outcomes	76	76
Target-comparator-neg controls	2,843,250* 76=216,087,000	769,476
Positive control outcomes	76* 3=228	228
Target-comparator-pos controls	2,843,250 * 228 = 648,261,00	662,484
Total comparisons	864,348,000	1,431,960





+ positive controls

	Theoretical	Observed (n>2,500)
Single ingredients	58	39
Single ingredient comparisons	58 * 57 = 3,306	1,296
Single drug classes	15	13
Single class comparisons	15 * 14 = 210	156
Dual ingredients	58 * 57 / 2 = 1,653	58
Single vs duo drug comparisons	58 * 1,653 = 95,874	3,810
Dual classes	15 * 14 / 2 = 105	32
Single vs duo class comparisons	15 * 105 = 1,575	832
Duo vs duo drug comparisons	1,653 * 1,652 = 2,730,756	2,784
Duo vs duo class comparisons	105 * 104 = 10,920	992
Total comparisons	2,843,250	10,278
Outcomes of interest	58	58
Target-comparator-outcomes	2,843,250 * 58 = 164,908,500	587,020
Negative control outcomes	76	76
Target-comparator-neg controls	2,843,250* 76=216,087,000	769,476
Positive control outcomes	76* 3=228	228
Target-comparator-pos controls	2,843,250 * 228=648,261,00	662,484
Total comparisons	864,348,000	1,431,960
Total	864,348,000 * 9= 7,779,132,000	1,431,960 * 9=12,887,640





US Insurance databases

- IBM[®] MarketScan[®] CCAE
- IBM[®] MarketScan[®] MDCD
- IBM[®] MarketScan[®] MDCR
- Optum© Clinformatics[®]

Japanese insurance databases

• Japan Medical Data Center

Korean national insurance databases

NHIS-NSC

US EHR databases

- Columbia University Medical Center
- Optum© PANTHER[®]

German EHR databases

• QuintilesIMS Disease Analyzer (DA) Germany

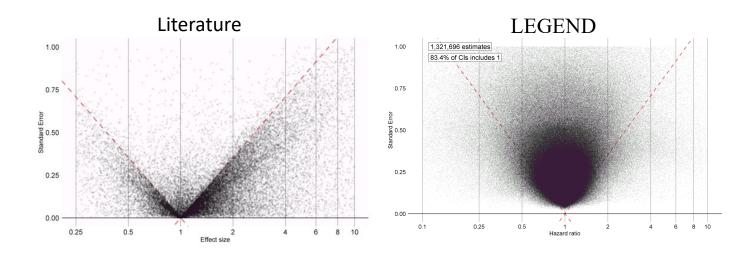
LEGEND knowledge base for hypertension



Head-to-head HTN drug comparisons Trials: 40 Comparisons: 10, 278

How does LEGEND perform?





- Best-practices **systematic design**, **evaluation** and empirical **calibration** return near nominal performance
- Provide a more complete and reliable evidence basis





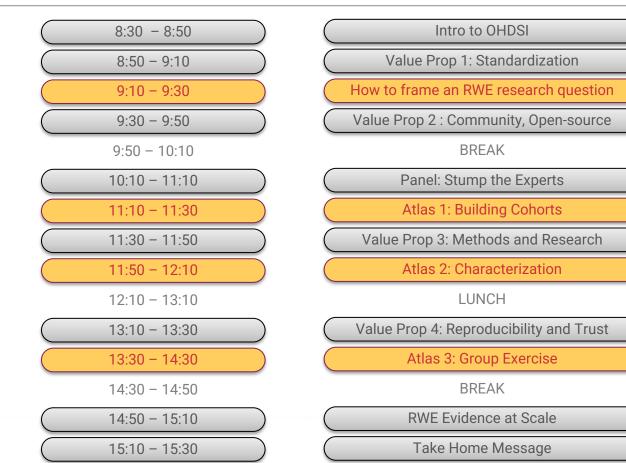
OHDSI has created the know how, people and the technical stack to make evidence generation an industrial process





Agenda





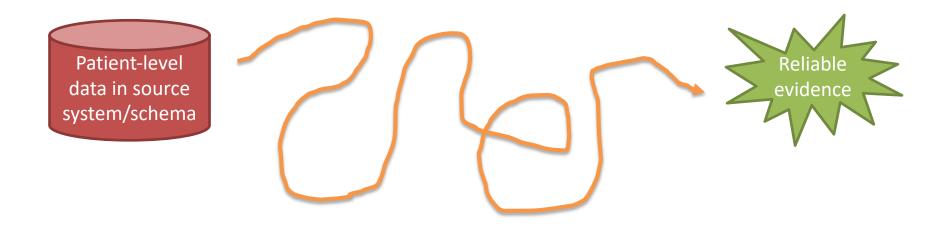


Evidence Generation

ME.250.961 Large Scale Observational Research Preparation

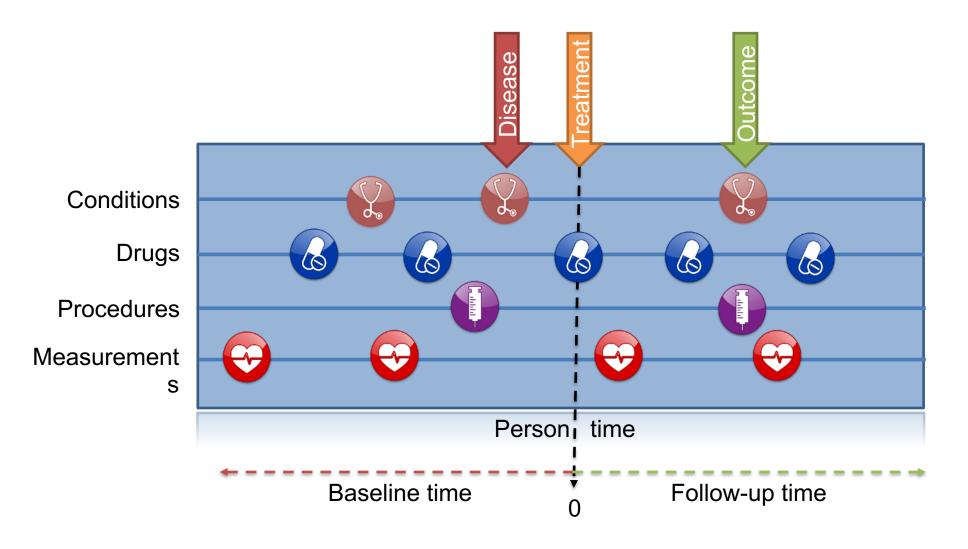


The journey to real-world evidence



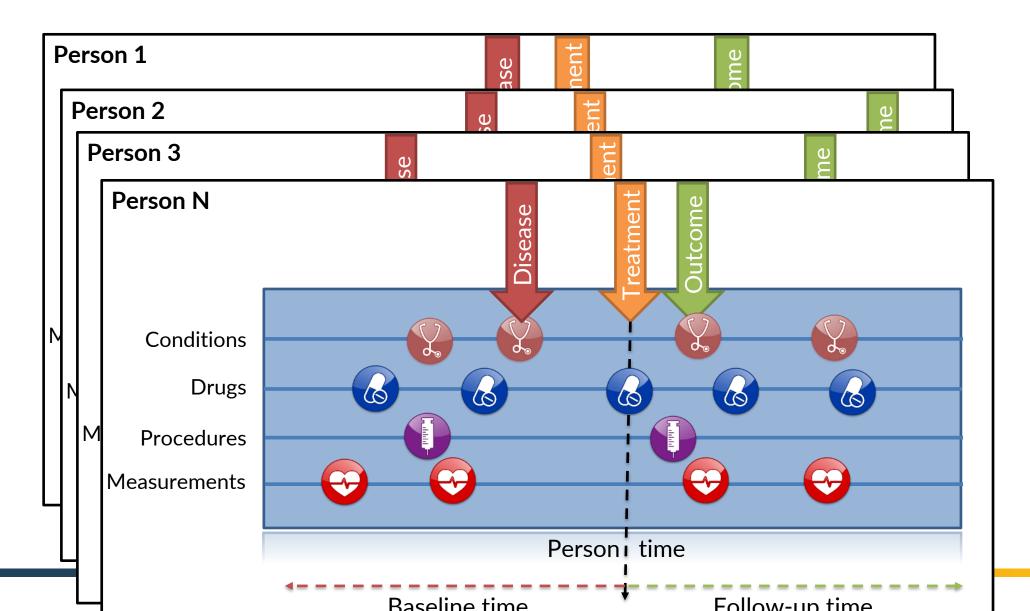


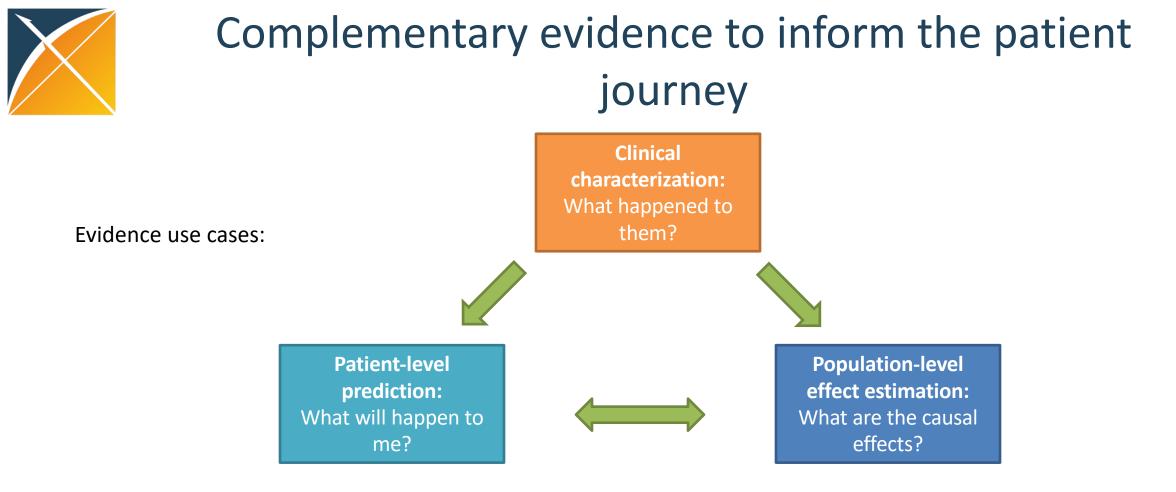
A Caricature of The Patient Journey





Each Observational Database Is Just an (Incomplete) Compilation of Patient Journeys





Analytic use case	Туре	Structure	Example
	Disease Natural History	Amongst patients who are diagnosed with <insert disease="" favorite="" your="">, what are the patient's characteristics from their medical history?</insert>	Amongst patients with rheumatoid arthritis , what are their demographics (age, gender), prior conditions, medications, and health service utilization behaviors?
Clinical characterization	Treatment utilization	Amongst patients who have <insert disease="" favorite="" your="">, which treatments were patients exposed to amongst <list of<br="">treatments for disease> and in which sequence?</list></insert>	Amongst patients with depression , which treatments were patients exposed to SSRI , SNRI , TCA , bupropion , esketamine and in which sequence?
	Outcome incidence	Amongst patients who are new users of <insert favorite<br="" your="">drug>, how many patients experienced <insert favorite<br="" your="">known adverse event from the drug profile> within <time horizon following exposure start>?</time </insert></insert>	Amongst patients who are new users of methylphenidate , how many patients experienced psychosis within 1 year of initiating treatment?
	Disease onset and progression	For a given patient who is diagnosed with <insert b="" favorite<="" your=""> disease>, what is the probability that they will go on to have <another complication="" disease="" or="" related=""></another> within <time b="" horizon<=""> from diagnosis>?</time></insert>	For a given patient who is newly diagnosed with atrial fibrillation , what is the probability that they will go onto to have ischemic stroke in next 3 years ?
Patient level prediction	Treatment response	For a given patient who is a new user of <insert favorite<br="" your="">chronically-used drug>, what is the probability that they will <insert desired="" effect=""> in <time window="">?</time></insert></insert>	For a given patient with T2DM who start on metformin , what is the probability that they will maintain HbA1C<6.5% after 3 years?
	Treatment safety	For a given patient who is a new user of <insert favorite<br="" your="">drug>, what is the probability that they will experience <insert adverse event > within <time exposure="" following="" horizon="">?</time></insert </insert>	For a given patients who is a new user of warfarin , what is the probability that they will have GI bleed in 1 year ?
Population-level	Safety surveillance	Does exposure to <insert drug="" favorite="" your=""> increase the risk of experiencing <insert adverse="" an="" event=""> within <time exposure="" following="" horizon="" start="">?</time></insert></insert>	Does exposure to ACE inhibitor increase the risk of experiencing Angioedema within 1 month after exposure start?
effect estimation	Comparative effectiveness	Does exposure to <insert drug="" favorite="" your=""> have a different risk of experiencing <insert (safety="" any="" benefit)="" or="" outcome=""> within <time exposure="" following="" horizon="" start="">, relative to <insert comparator="" treatment="" your="">?</insert></time></insert></insert>	Does exposure to ACE inhibitor have a different risk of experiencing acute myocardial infarction while on treatment , relative to thiazide diuretic ?



How does the rate of side effects / local problems (including secondary / palliative treatments needed) compare between treatments (open, laparoscopic, robot surgery, with or without lymph node dissection; brachytherapy, different forms of external beam radiation therapy), and which patient specific factors are associated with these adverse secondary endpoints?



 How does the rate of side effects / local problems (including secondary / palliative treatments needed) compare between
 treatments (open, laparoscopic, robot surgery, with or without lymph node
 dissection; brachytherapy, different forms
 of external beam radiation therapy) and

which patient specific factors are associated with these adverse secondary endpoints? <u>Characterization study: incidence rate</u>

Amongst patients with **prostate cancer receiving different treatments**, how many patients experienced **side effects/local problems** within <time horizon >?



 How does the rate of side effects / local problems (including secondary / palliative treatments needed) compare between
 treatments (open, laparoscopic, robot surgery, with or without lymph node
 dissection; brachytherapy, different forms
 of external beam radiation therapy) and

which patient specific factors are associated with these adverse secondary endpoints? <u>Population level estimation: comparative</u> <u>effectiveness</u>

Comparative effectiveness: Does exposure to **treatment A** have a different risk of experiencing **side effects/local problems** within <time horizon > , relative to **treatment B**?



- How does the rate of side effects / local problems (including secondary / palliative treatments needed) compare between treatments (open, laparoscopic, robot surgery, with or without lymph node dissection; brachytherapy, different forms of external beam radiation therapy), and which patient specific factors are associated with these adverse secondary endpoints?
- Characterization study: natural history

Amongst patients with **prostate cancer receiving different treatment A-Z**, what are the patient's characteristics from their medical history?



OMOP and OHDSI

What have we learned?

10-Nov-2023



To improve health

through community

and evidence





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.



20

Collaborators

OMOP Data by the Numbers

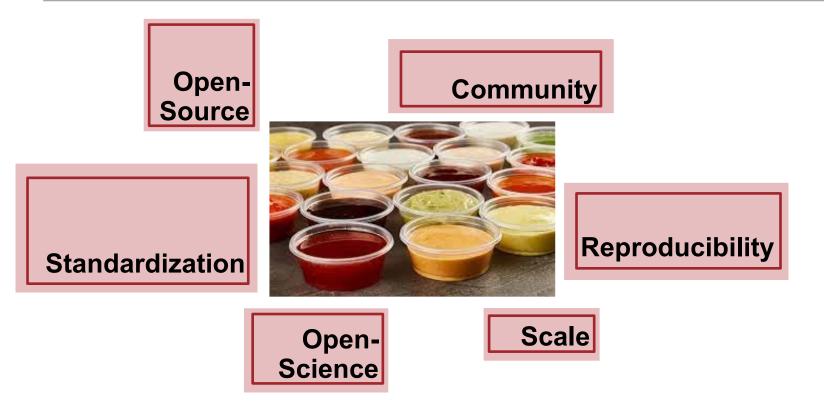
- 534 data sources
- 49 countries
- 956 million unique patient records
- approximately 12% of the world's population

OHDSI By The Numbers

- 3,758 collaborators
- 83 countries
- 21 time zones
- 6 continents
- 1 community

The Secret Sources





Data Standardization: OMOP CDM



INT INT INT

DATE

INT

INT

INT INT

DATE DATE

INT

INT INT

INT

FLOAT DATE

INT INT

DATE

INT

INT

INT

INT

VARCHAR INT

DATETIME DATE DATETIME INT

		-
Concept	a	Prov
concept ld	INT	provi
concept name	VARCHAR	provi
domain_id	VARCHAR	npi
vocabulary_id	VARCMAR	dea
concept_class_id	VARCMAR	speci
standard_concept	VARCHAR	care
concept_code	VARCHAR	Vear.
valid_start_date valid_end_date	DATE	provi
invalid_reason	VARCHAR	speci
		-< speci
Concept_class	<u> </u>	gend
concept class id	VARCHAR	gend
concept_class_name	VARCHAR	Care
<pre><concept_class_concept_id< pre=""></concept_class_concept_id<></pre>	INT	care
Vocabulary	1	care
vocabulary id	VARCHAR	place
vocabulary_name	VARCHAR	locat
vocabulary_reference	VARCHAR	care_
vocabulary_version	VARCHAR	place
<pre>vocabulary_concept_id</pre>	INT	Loca
Source to concept map	-	
source_code	VARCHAR	addr
source_concept_id	INT	addri
source_vocabulary_id	VARCHAR >	ctv
source_code_description	VARCHAR	state
<pre> target_concept_id </pre>	INT	zip
target_vocabulary_id	VARCHAR	count
valid_start_date	DATE	locat
valid_end_date invalid_reason	DATE	coun
Invalio_reason	VARLINAR	latitu
Domain	11	longi
domain id	VARCHAR	
domain_name	VARCHAR	
domain_concept_id	INT	Met
Concept Internet		meta
Concept_synonym	INT	meta
concept_synonym_name	VARCHAR	meta
<a>language concept id	INI	value
Cancestor concept ld	INT	value
<pre>descendant_concept_id</pre>	INT	value
min_levels_of_separation	INT	meta
max levels of separation	INT	meta
Concept_relationship	m	Cdm
Concept Id 1	INI	cdm_
Concept Id 2	INT	cdm_
relationship_id	VARCHAR	cdm.
valid_start_date	DATE	source
valid_end_date	DATE	source
invalid_reason	VARCHAR	cdm_
Relationship	11	sourc
relationship id	VARCHAR +	cdm_
relationship name	VARCHAR	cdm_
is hierarchical	VARCHAR	vocal
defines ancestry	VARCHAR	
reverse_relationship_id	VARCHAR >	
<prelationship_concept_id< pre=""></prelationship_concept_id<>	INT	
Drug_strength	-	
drug concept id	INT	
Ingredient concept id	INT	
amount value	FLOAT	
<amount concept="" id<="" td="" unit=""><td>INT</td><td></td></amount>	INT	
numerator_value	FLOAT	
<pre>fnumerator_unit_concept_id</pre>	INT	
denominator_value	FLOAT	
denominator_unit_concept_id	INT	
box size valid start date	INT	
valid_end_date	DATE	
invalid_reason	VARCHAR	
-		

	<u>v</u>
ovider_id	INT
rovider_name	VARCHAR
pl	VARCHAR
ea pecialty concept id	VARCHAR
are site id	INT
ear of birth	INT
ender_concept_id	INT
rovider_source_value	VARCHAR
pecialty_source_value	VARCHAR
pecialty source_concept_id	INT
ender_source_value ender_source_concept_id	VARCHAR
ender_source_concept_id	
are_site	Ś
are site id	INT
are site name	VARCMAR
lace of service concept id	INT
ocation_id	INT
are_site_source_value	VARCHAR
lace_of_service_source_value	VARCMAR
	0
ocation	Q
cation id	INT
ddress_1	VARCHAR
ddress_2	VARCHAR
ity rate	VARCHAR
ip	VARCHAR
ounty	VARCHAR
ocation_source_value	VARCHAR
ountry_concept_id	INT
ountry_source_value	VARCHAR
ititude	FLOAT
ongitude	FLOAT
Aetadata	
	INT
<u>setadata id</u> setadata_concept_id	INT
netadata_concept_id netadata_type_concept_id	INT
setadata_type_concept_id arte	VARCHAR
alue_as_string	VARCHAR
alue_as_concept_id	INT
alue_as_number	FLOAT
netadata_date	DATE
netadata_datetime	DATETIME
Maccomposite point	_
dm_source	
dm_source_name	VARCHAR
dm_source_abbreviation	VARCHAR
dm_holder	VARCHAR
ource_description	VARCHAR
ource_documentation_referenc	
dm_eti_reference	VARCHAR
ource_release_date	DATE
dm_release_date	DATE
dm_version dm_version_concept_id	VARCHAR
orabulary_version	VARCHAR
ocationary_version	VARCHAR

		_	_			7	,	
						f	٢	Party of the local division of the local div
4	INT					I	Н	Condition occurrence
2 toncept_id	INT					I	k	condition occurrence id
birth	INT					I		<pre>condition_concept_id</pre>
f_birth	INT					I	Н	condition_start_date
sirth	INT					I	L	condition_start_datetime
setime	DATETIME					I	L	condition_end_date
cept_id _concept_id id	INT					I	L	condition_end_datetime condition_type_concept_id
_concept_to	INT	>	_			I		condition_type_concept_lo
Jd	INT					I	L	stop_reason
له	INT	≻	1 I			ł	1	<pre>provider_id</pre>
ource_value	VARCHAR					I	Н	visit_occurrence_id
ource_value	VARCHAR					I	L	visit_detail_id
iource_concept_ld	INT					I	L	condition_source_value
rce_value rce_concept_ld	VARCHAR					I	Н	condition source concept condition status source vi
rce_concept_id	VARCHAR					I	Н	condition_status_source_vi
source value source concept ld	INT					I	L	Drug exposure
						I	L	drug exposure id
ation_period						I	Ł	Eperson_id
ion period id	INT					I	L	<a>dnug_concept_id
đ	INT					I	Н	drug_exposure_start_date
ion_period_start_date						I	Н	drug_exposure_start_datet
ion_period_end_date	DATE					I	L	drug_exposure_end_date
/pe_concept_id	INT	۰.				I	L	drug_exposure_end_datetr
	-					I	L	verbatim_end_date drug_type_concept_id
d	INT					I	L	stop_reason
ate	DATE					I	L	retilis
atetime	DATETIME					I	L	quantity
pe_concept_id	INT					I	L	days_supply
oncept_id	INT					I	L	sig
surce_value	VARCHAR					I	Н	<pre>route_concept_id</pre>
surce_concept_id	INT					l	r	fot number provider id
		6				ľ	T	visit occurrence id
courrence		1	L	L.		I	Н	visit_detail_id
urrence id d	INT	т	ľ	٦		I	L	drug_source_value
a cept_id	INT					I	L	drug source concept id
t_date	DATE					I	L	route source value
t_datetime	DATETIME					I	L	dose unit source value
date	DATE					I	Н	
datetime	DATETIME					I	L	Procedure_occurrence
_concept_id _id	INT					I	L	procedure occurrence id
jid	INT	ς.				I	٢	person_id
_id rce_value	INT	2	11			I	Н	<pre>procedure_concept_id procedure_date</pre>
rce_value rce_concept_id	INT					I	Н	procedure_datetime
_from_concept_id	INT					I	L	procedure_end_date
from_source_value	VARCHAR					I	L	procedure_end_datetime
ed to concept id	INT					I	L	<pre>procedure_type_concept_i</pre>
ed_to_source_value	VARCHAR					I	L	modifier_concept_id
visit_occurrence_id	INT	×	М	Μ		L	L	quantity
						ľ	T	<pre>provider_id</pre>
etail	, a		L	U	L	I	L	visit_occurrence_id visit_detail_id
<u>a u</u>	INT	2	ľ			11	L	procedure_source_value
a ail_concept_id	INT					II	Н	<pre>procedure_source_concept</pre>
al_start_date	DATE					Ш	L	modifier_source_value
al_start_datetime	DATETIME					Ш	L	
al end date	DATE					II	L	Device_exposure
sil_end_date sil_end_datetime	DATETIME					Ш	L	device exposure id
ail_type_concept_id	INT					II	v	erson_id
jid	INT					Ш		<pre>device_concept_id</pre>
_id sil_source_value	INT	1	1			Ш		device_exposure_start_dat device_exposure_start_dat
al_source_value all_source_concept_id	VARCHAR					Ш		device_exposure_end_date
_from_concept_id	INT					II		device_exposure_end_date
from source value	VARCHAR		r I	H		I		device_type_concept_id
_from_source_value id_to_source_value	VARCHAR		r I	H		I		unique_device_id
hi to concept id	INT		(I	U		I		production_id
_visit_detail_id	INT	2	M	M	1	I		quantity
isit_detail_id	INT	*	M	Ľ	۲	ľ	-	Fprovider_id
urrence_id	INT	2	ľ	1				visit_occurrence_id visit_detail_id
		1	1	H				visit_detail_id device_source_value
		_	λ					device_source_concept_id
		-						 unit_concept_id
								unit_source_value

			_		
_		_	-	- 16	\
					Measurement
	INT				measurement id
	INT			11	eperson_id
	INT				measurement_concept_id
	DATE				measurement_date
	DATETIME				measurement_datetime measurement_time
	DATETIME				measurement_type_concept_ic
	INT				operator_concept_id
ld .	INT				value_as_number
	VARCMAR				<pre>value_as_concept_id</pre>
	INT				unit_concept_id
	INT	>	-		range_low
	INT	>	ı I		range_high
	VARCMAR				<pre>provider_id</pre>
ld alue	INT				visit_occurrence_id
aue	VARCHAR				visit_detail_id
	6				measurement_source_value measurement_source_concept
	INT				unit_source_value
	INT				-unit_source_concept_id
	INT				value source value
	DATE				measurement event id
eme	DATETIME				meas event field concept id
	DATE				
me	DATETIME				Observation
	DATE				observation id
	INT				Eperson_id
	VARCHAR				<pre>cobservation_concept_id</pre>
	FLOAT				observation_date observation_datetime
	INT				<pre>conservation_type_concept_id</pre>
	VARCHAR				value_as_number
	INT				value_as_string
	VARCHAR				value_as_concept_id
	INT				
	INT	Z	Μ		<pre>unit_concept_id</pre>
	INT	-		111	Fprovider_id
	VARCHAR				visit_occurrence_id
	VARCMAR				visit_detail_id
	VARCHAR				observation_source_value
	THEFT				unit_source_value
					qualifier_source_value
_	INT	1			value source value
	INT				observation event ld
	INT				<pre>obs_event_field_concept_id</pre>
	DATE				Provide statements and st
	DATETIME				Note
	DATE				note id
	DATETIME				<pre>✓person_id</pre>
d	INT				note_date
	INT				note_datetime
	INT				note_class_concept_id
	INT	4	Ы		note_title
	INT	5			note_text
	VARCHAR				encoding_concept_id
Ud.	INT		H		language_concept_id
	VARCHAR				provider_id
-					visit_occurrence_id
					visit detail id
	INT				note_source_value
	INT				note_event_id
	INT				<pre>mote_event_field_concept_id</pre>
2	DATE				Note_nip
etime	DATETIME				
a de la compañía de	DATE				note_id
	INT				section_concept_id
	VARCHAR		H		snippet
	VARCHAR		H		offset
	INT				lexical_variant
	INT		H		<pre>mote_nlp_concept_id</pre>
	INT	2	Μ		<pre>id=_nip_source_concept_id</pre>
	INT	>	1		nip_system
	VARCHAR				nip_date
	INT		I I		nlp_datetime term_exists
	VARCHAR		H		term_exists term_temporal
	INT				term_modifiers
	and .		1		send_mouners

	6			Specimen
	INT		l	specimen id person_id
				person_id
1	INT			specimen_concept_id
	DATE			<pre>specimen_type_concept_id specimen_date</pre>
	VARCHAR			specimen_datetime
	VARCHAR			specimen_oatecime quantity
ept_id	INT			
				unit_concept_id
	FLOAT			anatomic_site_concept_id
	INT			disease_status_concept_id
	INT			specimen_source_id
	FLOAT			specimen_source_value
	FLOAT			unit_source_value
	INT			anatomic_site_source_value
	INT	2		disease status source value
	INT	1		Provide State of Stat
ue	VARCMAR			Fact_relationship
ncept_ld	INT			domain_concept_id_1
	VARCMAR			fact_id_1
	INT			
	VARCHAR			fact_id_2
	INT			relationship_concept_id
t id	INT			A State of the second state of the
				-
				Cost
	INT			cost id
	INT			cost_event_id
	INT			cost_domain_id
	DATE			<pre>cost_type_concept_id</pre>
	DATETIME			<pre>currency_concept_id</pre>
e, id	INT			total_charge
	FLOAT			total_cost
	VARCHAR			total_paid
	INT			paid_by_payer
	INT			paid_by_patient
	INT			paid_patient_copay
	INT			paid_patient_coinsurance
	INT	1	ч I	paid_patient_deductible
	INT	>		paid_by_primary
	VARCHAR	£ 1		paid_ingredient_cost
ept_id	INT			paid_dispensing_fee
	VARCHAR			payer_plan_period_id
	VARCHAR			amount_allowed
	VARCHAR			<pre>revenue_code_concept_id</pre>
	INT			revenue_code_source_value
,id	INT			drg_concept_id
-				drg_source_value
	INT	3	M-	Payer_plan_period
	INT			
	DATE			payer plan period id
	DATETIME			Person_id
	INT			payer_plan_period_start_date
	INT			payer_plan_period_end_date
	VARCHAR			<pre>payer_concept_id</pre>
	VARCHAR			payer_source_value
	INT		11	payer_source_concept_id
	INT			plan_concept_id
	INT			plan_source_value
	INT	~		plan_source_concept_id
	INT	5	11	<pre>sponsor_concept_id</pre>
		-		sponsor_source_value
	VARCMAR			<pre>sponsor_source_concept_id</pre>
				family_source_value
Ud	INT			stop_reason_concept_id
				stop_reason_source_value
				stop_reason_source_concept_id
	INT			
	INT	7	1	
	INT			
	VARCHAR			
	VARCHAR			
	VARCHAR			
	INT			
UK .	INT			

VARCHAR

DATE DATETIME VARCHAR VARCHAR VARCHAR

	Condition_era
r	condition era id
r	person_id
	condition_concept_id
	condition_era_start_date
TE	condition_era_end_date
TETIME	condition_occurrence_count
TAC	
1	Drug_era
r	drug era id
RCMAR	person_id
RCMAR	drug_concept_id
	drug_era_start_date
RCMAR	drug_era_end_date
RCMAR	drug_exposure_count
RCMAR	gap_days
9	Provide State Stat
	Dose era
	dose era id
_	person_id
	drug_concept_id
	unit_concept_id
	dose_value
	dose_era_start_date
	dose_era_end_date
r	Episode
r	episode id
RCHAR	
r	person_id
r	episode_concept_id
DAT	episode_start_date
DAT	episode_start_datetime
TAC	episode_end_date
	episode end datetime
TAC	episode_parent_id
TAC	episode_number
DAT	
DAT	episode_object_concept_id
DAT	episode_type_concept_id
DAT	episode_source_value
DAT	episode_source_concept_id
DAT	
nar r	Episode_event
DAT	episode_id
1	event_ld
RCHAR	episode event field concept id
r	
RCHAR	Cohort
	cohort_definition_id
- C.	subject_id
	cohort_start_date
r	cohort_end_date
TE	
TE	Cohort_definition
r	
	cohort_definition_id
ACHAR	cohort_definition_name
	cohort_definition_description
r	definition_type_concept_id
ACHAR	cohort_definition_syntax
	subject_concept_id
	cohort initiation date
RCHAR	
r	
RCHAR	
r	
RCHAR	
r	

Condition

-

INT

FLC

-

INT

04

VA



OHDSI Scales

	Theoretical	Observed (n>2,500)
Single ingredients	58	39
Single ingredient comparisons	58 * 57 = 3,306	1,296
Single drug classes	15	13
Single class comparisons	15 * 14 = 210	156
Dual ingredients	58 * 57 / 2 = 1,653	58
Single vs duo drug comparisons	58 * 1,653 = 95,874	3,810
Dual classes	15 * 14 / 2 = 105	32
Single vs duo class comparisons	15 * 105 = 1,575	832
Duo vs duo drug comparisons	1,653 * 1,652 = 2,730,756	2,784
Duo vs duo class comparisons	105 * 104 = 10,920	992
Total comparisons	2,843,250	10,278
Outcomes of interest	58	58
Target-comparator-outcomes	2,843,250 * 58 = 164,908,500	587,020
Negative control outcomes	76	76
Target-comparator-neg controls	2,843,250* 76=216,087,000	769,476
Positive control outcomes	76* 3=228	228
Target-comparator-pos controls	2,843,250 * 228=648,261,00	662,484
Total comparisons	864,348,000	1,431,960
Total	864,348,000 * 9= 7,779,132,000	1,431,960 * 9=12,887,640





US Insurance databases

- IBM[®] MarketScan[®] CCAE
- IBM[®] MarketScan[®] MDCD
- IBM[®] MarketScan[®] MDCR
- Optum© Clinformatics[®]

Japanese insurance databases

• Japan Medical Data Center

Korean national insurance databases

NHIS-NSC

US EHR databases

- Columbia University Medical Center
- Optum© PANTHER®

German EHR databases

• QuintilesIMS Disease Analyzer (DA) Germany

Open Source Community



262 Repositories 30 M+ lines of code 681 Developers 31 organizations 47,672 commits 2,838 GitHub Forks 4,168 GitHub Stars 5,547 GitHub Subscribers

				Open	pull-		
Package	Version	Maintainer(s)	Availability	issues	requests	Build status	Coverage
Achilles	v1.7.2	Frank DeFalco	CRAN	29	3	C R check passing	\varTheta codecov
Andromeda	v0.6.3	Adam Black	CRAN	13	2	C R check passing	🜳 codecov 🧧
BigKnn	v1.0.2	Martijn Schuemie	GitHub	0	0	R check passing	Codecov 9
<u>BrokenAdaptiveRidge</u>	v1.0.0	Marc Suchard	CRAN	2	0	R check passing	Codecov 9
<u>Capr</u>	v2.0.7	Martin Lavallee	GitHub	2	0	C R check passing	Codecov 8
Characterization	v0.1.2	Jenna Reps	GitHub	11	1	C R check failing	Codecov unkn
<u>CirceR</u>	v1.3.1	Chris Knoll	GitHub	3	1	C R check passing	Codecov 8
<u>CohortDiagnostics</u>	v3.2.4	Jamie Gilbert	GitHub	56	2	C R check passing	Codecov 9
CohortExplorer	v0.1.0	Gowtham Rao	CRAN	0	0	C R check passing	Codecov 10
CohortGenerator	v0.8.1	Anthony Sena	GitHub	19	2	C R check passing	Codecov 9
<u>CohortMethod</u>	v5.1.0	Martijn Schuemie	GitHub	14	1	C R check failing	Codecov 8
<u>Cyclops</u>	v3.3.1	Marc Suchard	CRAN	18	0	C R check failing	🔶 codecov 🧧
DatabaseConnector	v6.2.4	Martijn Schuemie	CRAN	12	0	C R check passing	🔶 codecov 🧧
DataQualityDashboard	v2.4.1	Katy Sadowksi	GitHub	43	7	() R check passing	Codecov 8
DeepPatientLevelPrediction	v2.0.0	Egill Fridgeirsson	GitHub	18	1	C R check failing	Codecov 10
EmpiricalCalibration	v3.1.1	Martijn Schuemie	CRAN	1	0	C R check passing	🔶 codecov 🔒
EnsemblePatientLevelPrediction	v1.0.2	Jenna Reps	GitHub	5	0	C R check passing	🖓 codecov 🛛 unka
Eunomia	v1.0.2	Frank DeFalco	GitHub	10	1	C R check passing	Codecov 7
EvidenceSynthesis	v0.5.0	Martijn Schuemie	CRAN	3	0	C R check passing	Codecov 7
FeatureExtraction	v3.3.1	Anthony Sena	GitHub	44	6	C R check failing	🔶 codecov 🧕
<u>Hydra</u>	v0.4.0	Anthony Sena	GitHub	6	7	C R check failing	Codecov 8
IterativeHardThresholding	v1.0.2	Marc Suchard	CRAN	1	0	C R check passing	Codecov 9
MethodEvaluation	v2.3.0	Martijn Schuemie	GitHub	1	0	C R check passing	🔶 codecov 🧕
OhdsiSharing	v0.2.2	Lee Evans	GitHub	0	1	C R check passing	🔶 codecov 🧧
<u>OhdsiShinyModules</u>	v2.0.0	Jenna Reps	GitHub	108	2	C R check failing	👇 codecov 💈
ParallelLogger	v3.3.0	Martijn Schuemie	CRAN	4	0	C R check passing	🔶 codecov 🔒
PatientLevelPrediction	v6.3.6	Jenna Reps & Pete Rijnbeek	er GitHub	47	0	R check failing	P codecov 8

OHDSI Methods and Research



Three ideas

2. Objective Diagnostics

Engineering open science systems that build trust into the real-world evidence generation and dissemination process

3. Standardized software



Open-source software



Perspective LEGEND answers large sets of related questions at once (eg, the effects of many treatments for a disease Principles of Large-scale Evidence Generation and on many outcomes). Aim: Avoids publication bias, Evaluation across a Network of Databases (LEGEND) achieves commehensiveness of results, and allows for Martijn J. Schuemie (3^{1,2}, Patrick B. Ryan^{1,3}, Nicole Pratt⁴, RuiJun Chen (3^{1,5}, Seng Chan You⁶, Harlan M. Krumholz⁷, David Madigan⁶, George Hripcsak^{2,6}, and an evaluation of the overall coherence and consistency Marc A. Suchard^{2,1}

2. Dissemination of the evidence will not depend on the estimated effects. All generated evidence is disseminated at once. All mit works within this and enhances transparent

1. LEGEND Principles

LEGEND in Principle

LEGEND (Large-scale Evidence Generation and Evaluation across a Network of Databases) applies high-level analytics to perform observational research on hundreds

of millions of patient records within OHDSI's international database network. LEGEND is based on 10 guiding principles that were published in JAMIA (August, 2020)

and are listed below.

of the generated evidence.

1. LEGEND will generate evidence at a

large scale. Instead of answering a single question

at a time (on the effect of 1 treatment on 1 outcome).

METHODS RESEARCH

3. LEGEND will generate evidence using a prespecified analysis design. All analyses, including the research questions that will be answered, will be decided prior to analysis execution. Aim: Avoids P hacking

4. LEGEND will generate evidence by consistently applying a systematic process across all research questions. This principle precludes modification of analyses to obtain a desired answer to any specific question. This does not imply a simple one-size-fits-all process, rather that the logic for modifying an analysis for specific research questions should be explicated and applied systematically. Aim: Avoids P hacking and allows for the evaluation of the operating characteristics of this process (Principle 6)

5. LEGEND will generate evidence using best practices. LEGEND answers each question using current best practices, including advanced methods to address confounding, such as propensity scores. Specifically, we will not employ suboptimal methods (in terms of bias) to achieve better computational efficiency. Aim: Minimizes bias

6. LEGEND will include empirical evaluation through the use of control questions. Every LEGEND study includes control questions. Control questions are questions where the answer is known. These allow for measuring the operating characteristics of our systematic process including residual bias. We subsequently account for this observed residual bias in our P values, effect estimates and confidence intervals using empirical calibration. [7.8] Aim: Enhances transparency on the uncertainty due to residual bias

7. LEGEND will generate evidence using open-source software that is freely available to all. The analysis software is com to review and evaluation and is available for replicating analyses down to the smallest detail Aim; Enhances transparency and allows replication

8. LEGEND will not be used to evaluate new methods. Even though the same intrastructure used in LEGEND may also be used to evaluate new causal inference methods, generating clinical evidence should not be performed at the same time as method evaluation. This is a corollar of Principle 5, since a new method that still requires evaluation cannot already be best practice. Also, generating evidence with unproven methods can harmore the interpretability of the clinical results. Note that I EGEND does evaluate how well the methods it uses perform in the specific context of the questions and data used in a LEGEND study (Principle 6). Aim: Avoids bias and improves interpretability.

9. LEGEND will generate evidence across a network of multiple databases. Multiple heterogeneous databases (different data capture processes, health care systems, and populations) will be used to generate the evidence to allow an assessment of the replicability of findings across sites. Alm: Enhances generalizability and uncovers potential between-site heterogeneity.

10. LEGEND will maintain data confidentiality; patient-level data will not be shared between sites in the network Not sharing data will ensure patient privacy, and comply with local data governance rules. Aim: Privacy,

#JoinTheJourney	47	OHDSI.org
-----------------	----	-----------

Distributed data network, standardized to common data model Network coordination Data quality evaluation Phenotype development and evaluation Analysis reliability evaluation Fail System characteristics: Standardized procedures with defined inputs and outputs Analysis packages implementing scientific best practices consistently applied across all data partners, generating consistent

output for network synthesis Reproducible outputs generated by open-source analysis libraries

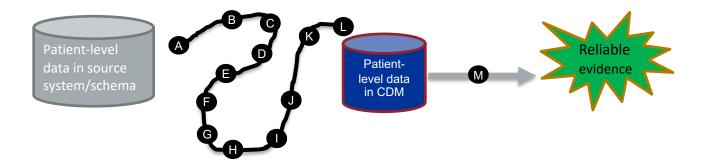
- developed and validated with verifiable unit-test coverage
- Pre-specified and objective decision thresholds for go/no go criteria
- Measurable operating characteristics of system performance



Reproducibility and Trust



We need to make studies repeatable, reproducible, replicable, generalisable, and robust



A Common Data Model enables standardised analytics to generate reliable evidence.

Where do you find us?



Home: https://www.ohdsi.org

Book of OHDSI: <u>https://book.ohdsi.org</u>

Methods and Tools: <u>https://github.com/OHDSI</u>

Common Data Model: https://ohdsi.github.io/CommonDataModel

Vocabularies: https://athena.ohdsi.org

Studies: https://github.com/OHDSI/ohdsistudies

Workgroups: <u>https://www.ohdsi.org/workgroups</u>



Join the Journey!

