Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Healthc (Amst). Author manuscript; available in PMC 2024 April 11.

Published in final edited form as:

Healthc (Amst). 2020 December ; 8(4): 100458. doi:10.1016/j.hjdsi.2020.100458.

The Longitudinal Epidemiologic Assessment of Diabetes Risk (LEADR): Unique 1.4 M patient Electronic Health Record cohort

Howard A. Fishbein[a,*], **Rebecca Jeffries Birch**[a], **Sunitha M. Mathew**[a], **Holly L. Sawyer**[a], **Gerald Pulver**[b], **Jennifer Poling**[c], **David Kaelber**[d], **Russell Mardon**[a], **Maurice C. Johnson**[a], **Wilson Pace**[e], **Keith D. Umbel**[a], **Xuanping Zhang**[f], **Karen R. Siegel**[f], **Giuseppina Imperatore**[f], **Sundar Shrestha**[f], **Krista Proia**[f], **Yiling Cheng**[f], **Kai McKeever Bullard**[f], **Edward W. Gregg**[f], **Deborah Rolka**[f], **Meda E. Pavkov**[f]

[a]Westat, Rockville, MD, USA

[b]University of Colorado Anschutz Medical Campus, Denver, CO, USA

[c]Cherokee Health Systems Inc, Knoxville, TN, USA

[d]The MetroHealth System and Case Western Reserve University, Cleveland, OH, USA

[e]DARTNet, Aurora, CO, USA

[f]Centers for Disease Control and Prevention, Division of Diabetes Translation, Atlanta, GA, USA

Abstract

Background: The Longitudinal Epidemiologic Assessment of Diabetes Risk (LEADR) study uses a novel Electronic Health Record (EHR) data approach as a tool to assess the epidemiology of known and new risk factors for type 2 diabetes mellitus (T2DM) and study how prevention interventions affect progression to and onset of T2DM. We created an electronic cohort of 1.4 million patients having had at least 4 encounters with a healthcare organization for at least 24-months; were aged ≥18 years in 2010; and had no diabetes (i.e., T1DM or T2DM) at cohort entry or in the 12 months following entry. EHR data came from patients at nine healthcare organizations across the U.S. between January 1, 2010–December 31, 2016.

Results: Approximately 5.9% of the LEADR cohort (82,922 patients) developed T2DM, providing opportunities to explore longitudinal clinical care, medication use, risk factor trajectories, and diagnoses for these patients, compared with patients similarly matched prior to disease onset.

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

[*]Corresponding author. Westat 1600 Research Blvd TC3030, Rockville, MD, 20850, USA. HowardFishbein@westat.com (H.A. Fishbein).

Financial disclosure

No financial disclosures were reported by the authors of this paper.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.hjdsi.2020.100458.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Conclusions: LEADR represents one of the largest EHR databases to have repurposed EHR data to examine patients' T2DM risk. This paper is first in a series demonstrating this novel approach to studying T2DM.

Implications: Chronic conditions that often take years to develop can be studied efficiently using EHR data in a retrospective design.

Level of evidence: While much is already known about T2DM risk, this EHR's cohort's 160 M data points for 1.4 M people over six years, provides opportunities to investigate new unique risk factors and evaluate research hypotheses where results could modify public health practice for preventing T2DM.

Keywords

Chronic disease; Diabetes mellitus; Epidemiologic methods; Epidemiologic research design; Big data; Electronic health records; Public health informatics; Public health practice

1. Introduction

Diabetes is the seventh-leading cause of death in the United States (U.S.), and is a large economic burden on the U.S. healthcare system – approximately $327 billion in 2017.[1] An estimated 12.2% of the U.S. adult population, or 30.2 million adults, has either diagnosed or undiagnosed diabetes, 95% of which is type 2 diabetes mellitus (T2DM).[2] A further 84 million Americans have prediabetes.[2] From other reported research, up to 70% of Americans with prediabetes will eventually develop T2DM.[3]

Electronic health records (EHRs) from healthcare organizations provide patient-level data that complement national survey data in building a complex picture of the epidemiology of T2DM risk.[4] The availability of longitudinal clinical data for a large and diverse EHR-based cohort provides opportunities to track progression to T2DM.[5] Patient demographics, medical history, physical examination, laboratory testing, and medication data from EHRs can help identify incident cases of diabetes and its subtypes. Reviews of the frequency and completeness of recommended tests, prescriptions, screening, and treatment goals can help assess the quality of diabetes care delivery. Analysis of family history, demographics, body mass index (BMI), smoking status, prescriptions, diagnoses, and procedures aids in identifying diabetes risk factors, microvascular and macrovascular complications, and comorbidities. While healthcare organizations are primarily focused on clinical treatments and patient outcomes, their databases are rich with information that can be repurposed to investigate epidemiologic patterns and prevention strategies. Despite these advantages, only a few studies, such as SUPREME-DM,[6] have focused on analyzing risk of developing T2DM using EHRs.

The goals of LEADR are to: build a robust, standardized database using EHRs; examine a patient's risk profile prior to the onset of elevated glycemia; identify risk factors for progression to elevated glycemia; examine changes in risk factors over time and corresponding changes in clinical outcomes; and create a dynamic risk profile for patients to assess their individual risk status for developing T2DM. This manuscript describes the architecture of LEADR and establishes this landmark database.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

2. Methods

2.1. Healthcare organizations

In creating an improved EHR data repository for research, we partnered with healthcare organizations including primary care and multispecialty integrated delivery systems with ambulatory and inpatient components located across the U.S. These clinical sites were aggregated to represent four regions: Northeast (Connecticut, Vermont, Ohio), South (North Carolina, Tennessee), Rocky (Wyoming, Nebraska, Colorado), and West (California, Idaho, Washington) (Figure S1). They included a large Accountable Care Organization, private practices, Federally Qualified Health Centers (FQHCs), university practices, and FQHC Look-Alikes (i.e., community-based health care providers meeting requirements of the Health Research Services Administration Health Center Program but do not receive Health Center Program funding).[7] The participating organizations provide care in inner-city, rural, and suburban areas, including underserved populations.

An innovation included our working collaboratively with health plans to create the unique EHR repository we were developing. We first identified data domains of interest. Domains included demographics, vital signs, medications, diagnoses, medical and surgical history, physical exam results, laboratory results, and referral history. The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) v4 was used to standardize the disparate coding schemes present in our partner EHRs, such as the International Classification of Diseases (ICD), National Drug Code, Current Procedural Terminology, and the Systematized Nomenclature of Medicine, into a common vocabulary. The application of OMOP CDM terminologies, vocabularies, and coding schemes allowed for the aggregation of disparate observational databases to a unified LEADR database.[8]

To maximize the EHR system for research purposes and maintain confidentiality, partnering LEADR healthcare organizations were masked, personally identifiable information was removed and coded identifications were created for patients. Date of birth was replaced with age of patient as of January 1, 2010. Healthcare service dates were masked by applying to the dates a randomly selected integer between −183 and + 183. This offset remained consistent across each patient's full set of data, thereby ensuring temporal relationships remained intact. Since the LEADR database did not contain Personal Identifying Information (PII), the study was exempt from Institutional Review Board (IRB) review following review and approval from the Colorado Multiple IRB.

LEADR outcomes could affect the delivery of clinical care and therefore data quality was of highest priority. To meet this study aim, a data quality workgroup reviewed the consistency, accuracy, and completeness of the data by following a conceptual framework for assessing EHR data quality[9]: 1) attribute domain rules—checking for data value anomalies (e.g., outliers, missingness, incorrect units) for individual variables; 2) relational integrity rules—comparing data elements in one table to related elements in other tables; 3) historical data rules—viewing temporal relationships to identify data gaps; 4) state-dependent object rules—checking for logical inconsistencies (e.g., prenatal ultrasounds should precede a pregnancy outcome); and 5) attribute dependency rules—examining

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

conditional dependencies based on clinical scenarios (e.g., women should not have a diagnosis of prostate cancer).

Following quality assessment, data were aggregated into four regional databases using OMOP CDM v4 relational data model (Figure S2). Six OMOP tables created included Person, Condition, Observation, Procedure, Drug and Visit tables. Fig. 1 shows the data flow for the LEADR cohort.

2.2. Constructing the LEADR cohort

In constructing the LEADR Cohort, the primary aim was to construct a database that 1) took advantage of the longitudinal data provided by DARTNet, 2) ensured each patient had sufficient data over an extended period of time, and 3) was optimized to assess new cases of diabetes and corresponding risk factors. Based on these objectives we identified a start date as the first appearance in the database for each patient, including a patient in the cohort given sufficient data were present to support analytic objectives. Patients included in the LEADR cohort met all five of the following inclusion and exclusion criteria:

1. Patients had at least four encounters for any health care service, and each encounter occurred at least 2 weeks apart.
2. There were at least 24 months between the date of the first encounter and the date of the last encounter.
3. Patients were ≥18 years old on January 1, 2010.
4. Patients were free of T1DM, T2DM, or unspecified diabetes at cohort entry and in their first 12 months in the cohort.
5. Patients who developed T1DM during the cohort period were excluded.

In determining the number of encounters required for inclusion in the LEADR cohort, there is an inherent tradeoff between inclusivity and measurement ability, and we wanted to maximize the usefulness of the EHR repository we were creating. Supplemental Table S1 presents the demographic distributions, prevalence of selected risk factors, and T2DM incident cases for cohorts requiring differing numbers of encounters. We chose four encounters with a healthcare organization as part of the criteria for cohort inclusion. Requiring a minimum number of encounters improves data availability for all participants over time and thus allows for examining risk factor trajectories.

We chose a 24-month interval between the first and last encounter to ensure sufficient time for assessing longitudinal changes in risk factors and diabetes status.

Our approach for identifying diabetes cases derived from case definitions from previous literature as well as feedback from clinicians and statisticians with EHR expertise. Methods for defining diabetes risk vary in complexity and in clinical utility, as well as the extent to which they have been validated. However, common themes were found from the varying empirical approaches.[6,10-13] For example, to ensure the LEADR cohort identifies new diabetes cases, a “washout period” is applied requiring no evidence of diabetes in the data within the first year of the patient’s EHR record.[6,11,14] In addition, algorithms are applied

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

to reasonably discriminate between type 1 and type 2 diabetes, applying a combination of specified claims, laboratory, and medication parameters.[15,16]

For prevalent cases, the diabetes definition criteria was inclusive to ensure we would capture and exclude all possible existing diabetes cases from the cohort. We excluded patients with a diagnosis of T1DM, T2DM, or unspecified diabetes, prescription for an antidiabetic agent, or laboratory results in the diabetes range at their first encounter or in the following 12 months. Those with hemoglobin HbA1c ≥ 6.5%, fasting plasma glucose (FPG) ≥126 mg/dL, or random blood glucose ≥200 mg/dL[17] were excluded. While one random glucose ≥200 mg/dL is used clinically in conjunction with symptoms to diagnose diabetes,[17] since symptom data are not included in LEADR, we relied on the random glucose measurements alone to ensure exclusion of all possible cases of prevalent diabetes. For pregnant women, laboratory values were not used to ascertain diabetes. A woman with a FPG ≥126 mg/dL during her pregnancy was not excluded, unless she also had any diabetes diagnosis or was prescribed an antidiabetic agent within 12 months of cohort entry.

2.3. Case and high risk factor definitions

Table 1 provides the LEADR definitions of diabetes and risk factors. Incident diabetes was captured with three types of data: diagnosis records reflecting ICD codes, prescription drugs, and laboratory results. The requirements for defining incident T2DM cases are different from those used to exclude prevalent cases from the cohort. To identify incident cases, we used a more restrictive diagnostic criteria to ensure we were identifying patients with a high likelihood of having diabetes. A new diabetes case required two separate diagnosis records of diabetes (T2DM or unspecified diabetes) that were at least 14 days apart. Patients with one diabetes diagnosis and a prescription for either metformin or glucagon-like peptide-1 (GLP-1) agonists qualified as a case. Patients with prescriptions of metformin or GLP-1 agonists without a diabetes diagnosis were not classified as a diabetes case because these drugs are also prescribed to individuals with prediabetes or conditions associated with insulin-resistance such as polycystic ovary syndrome.[18,19,20] Patients with a prescription for an antidiabetic agent (other than metformin and GLP-1 agonists) were counted as a diabetes case as were patients with an HbA1c result ≥6.5% or a fasting blood glucose result ≥126 mg/dL.

The utilization of random glucose measurements to identify incident diabetes cases is more complicated with EHR data since clinically a random glucose ≥200 mg/dL is used in combination with symptoms to diagnose diabetes.[17] Because the sensitivity and specificity of random glucose tests to diagnose diabetes are lower than for HbA1c tests,[21] and because LEADR does not have symptom data, we required two random blood glucose values of ≥200 mg/dL at least 14 days apart, or a single random glucose value ≥200 mg/dL along with one diabetes diagnosis, or a single random glucose ≥250 mg/dL to count as an incident diabetes case. When random blood glucose was the sole metric available for assessing diabetes, we selected a higher value (≥250 mg/dL) since such a high value would indicate the person very likely has diabetes. This value was based on diagnostic criteria for diabetic ketoacidosis.[21] Using this criteria to flag T2DM incident cases maximizes the use of available data and the likelihood that identified cases are true cases. The date of diabetes

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

diagnosis was set to the earliest of 1) diabetes diagnosis, 2) prescription for antidiabetic drug, or 3) diabetes-related laboratory results above a set threshold.

Prediabetes was defined using diagnosis records and laboratory measurements (Table 1). Random blood glucose results were not used for defining prediabetes, as this is not standard clinical practice.[22] Additionally, research on the efficacy of using random blood glucose to identify prediabetes as well as the thresholds to apply is mixed.[23-25]

We defined hypertension and elevated lipids using ICD diagnosis records, prescription drug records, and blood pressure measurements or laboratory results. Obesity was defined using ICD diagnosis records or body measurements. Depression was defined based on ICD diagnosis records. To determine history of tobacco use, we used records from the Condition table (e.g., diagnosis of tobacco dependence), Observation table (e.g., text denoting current use or quit smoking), Procedure table (e.g., treatment for tobacco use), and Drug table (e.g., prescriptions for nicotine replacement therapy). We defined family history of diabetes from the Observation and Diagnosis tables. Table 1 presents the details of these definitions.

Patients with no record of a specific diagnosis or a specific medication to treat that condition, or no related laboratory results/body measurements, were classified as missing for that particular health condition. A patient who did not meet the requirements of steps one to three in Table 1, and who had no HbA1c, fasting plasma glucose, or random glucose results, was classified as not able to assess their diabetes status rather than being diabetes free.

Data processing and statistical analysis performed in 2018 used SAS v9.4.[26]

3. Results

From the population of 3.7 million patients receiving healthcare services between January 1, 2010, and December 31, 2016, 1.4 million patients met the inclusion criteria and were included in the LEADR cohort. There were 34.8 million encounters in LEADR with a median of 18 encounters per patient. The mean number of days patients were in the cohort (days between the first and last encounter) was 1648 days, or 4.5 years. Table 2 presents the distributions of number of records per patient for various data types. The count of records for the data types sum to much higher than the total encounters because counts for each data type were not required to be 14 days apart. The percent of patients in the cohort with no records of a specific data type ranged from 2% for condition/diagnosis to 98% for FPG. Eighty-three percent of patients had no HbA1c measurements and 59% had no random glucose measurements.

Supplemental Table S2 presents the number and percent of incident T2DM cases by a hierarchical identification method. Forty-seven percent of the 82,922 incident cases were detected through ICD diagnosis records, 20% through medications, and 32% through laboratory measures.

Table 3 presents the cohort profile, including the demographic distribution and the number of T2DM incident cases. Sixty-one percent of the LEADR cohort is female. The cohort is

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

70.1% non-Hispanic white, 9.6% Hispanic, 8.5% non-Hispanic black, 1.6% Asian, 0.4% American Indian/Alaska Native, and 0.8% other race/ethnicity (Table 3). The mean age is 47.4 years and the age distribution is fairly even across six age categories ranging from 18 to 24 years–65 years and older. Over the cohort period, 82,922 patients (5.9% of the cohort) developed T2DM.

4. Discussion

We created a longitudinal cohort of 1.4 million adults with diverse demographic characteristics and serial measurements of multiple clinical indicators covering a 7-year period. While LEADR was not developed to be population-based, the geographic diversity of the healthcare organizations, and the types of organizations contributing data to LEADR, provide a robust look into a cohort of individuals and their risk for developing T2DM. With 82,922 patients developing T2DM over the 7-year study period, the LEADR cohort provides a foundation for investigating both common and relatively rare risk factors for T2DM, which we are pursuing.

While utilizing EHR data to study disease progression can be cost-efficient, assembling a database from multiple healthcare organizations is challenging. The methods used in constructing LEADR improve the approach to using EHR data repurposed to study disease and related outcomes. A strength of LEADR is the use of the OMOP CDM to standardize and harmonize diagnostic, procedural, medication, and laboratory coding across EHR systems, allowing for aggregation of data from different sources. Likewise, carefully constructed cohort inclusion criteria, well-delineated case definitions, and definitions of risk factors facilitate examination of diabetes diagnosis, disease progression, and their association with risk factors.

Limitations in using EHR data for epidemiologic research include limits in representativeness; data availability and completeness; availability of environmental, community, dietary and behavioral diabetes risk factors, and information on diabetes prevention approaches. Representativeness is constrained to those who obtain care from healthcare organizations contributing data.[27] LEADR results can best be generalized to populations that are seeking and receiving healthcare at similar healthcare organizations. Further, in the first year of a longitudinal cohort, despite the one-year washout period, diabetes diagnosis may still be present among some individuals due to the lack of historical health status data.[13]

We recognize that missing data can be of concern when analyzing EHR records. LEADR has two broad categories of missing data: 1) incomplete records where tests may have been performed but data were not recorded, and 2) systematic differences where there was an apparent missingness pattern. There are various ways to address missingness in analyses, and given the variety of LEADR studies planned, we do not see a one-size-fits-all solution. Future LEADR manuscripts will choose to use techniques for handling informative missing data (missing not at random) such as, imputation, pattern mixture models or inverse-probability weighting to address missing data. A rationale and justification for the specific approach in handling missing data will be detailed.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Clinical documentation in EHRs is often not highly structured, nor is it as complete and consistent as data collected primarily for research.[28,29] Further, EHR diagnoses may be nonspecific and not updated,[30] and environmental, community, and behavioral risk factors in diabetes disease progression are often lacking.[28,31]

Despite limitations, the LEADR cohort is providing a unique tool for a variety of studies. For example, 1) exploration and validation of known and new risk factors and their associations with progression to T2DM; 2) investigation of the relationship between patient-level characteristics and the assignment or lack of assignment of prediabetes diagnosis by clinical providers; 3) examination of the longitudinal relationship between traditional and emerging behavioral and metabolic risk factors and the progression to prediabetes, T2DM, and diabetes-related morbidity; 4) estimation of the levels of receipt and effectiveness of T2DM prevention programs, such as the National Diabetes Prevention Program, and including other nutritional and lifestyle counseling programs, and the use of preventive medications and interactions with other prescribed medications; 5) study of factors that affect progression to T2DM and diabetes complications; 6) examination of the long-term impact of T2DM preventive behaviors and services on vascular, neuropathic, and aging-related outcomes; and 7) development and validation of risk equations/health profile to predict the likelihood of progression to T2DM and diabetes-related complications. We are also linking patient records to community characteristics, such as income, education, unemployment, and urban/rural data to examine the impact of socioeconomic factors and social determinants of health associated with risk of developing T2DM.

5. Conclusions

LEADR established a unique, diverse, longitudinal electronic cohort of 1.4 million patients that serves as an improved approach to use EHR data as a resource to assess the role of traditional and emerging risk factors and preventive services in the progression to T2DM and related morbidity. The availability of a large and diverse body of patient data, spanning anthropometry, and metabolic markers, to patterns of care utilization, will allow the development, testing, and validation of alternative risk stratification approaches for the screening, testing, and referral of persons likely to benefit from T2DM prevention services.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This project was solely funded by the Centers for Disease Control and Prevention under contract HHSD200201587699 to Westat Inc. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention or the National Institutes of Health. The article contents have not been presented elsewhere. The authors declare that they have no conflict of interest. The authors of this paper reported no financial disclosures.

Study concept, data collection, analysis, and manuscript preparation and review completed by Howard A. Fishbein, DrPH, Rebecca Jeffries Birch, MPH, Sunitha M. Mathew, MS, Holly L. Sawyer, BA, Gerald Pulver, PhD, Jennifer Poling, MBA, and David Kaelber, MD, PhD, MPH.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Study concept, data collection and analysis completed by Russell Mardon, PhD, Maurice C. Johnson, MPH, and Wilson Pace, MD. Programming support, data collection and harmonization provided by Keith Umbel, BS. Study concept, analysis input and manuscript review provided by Xuanping Zhang, PhD, Karen R. Siegel, PhD, Giuseppina Imperatore, MD, PhD, Sundar Shrestha, PhD, Krista Proia, MPH, Yiling Cheng, PhD, Kai McKeever Bullard, PhD, Edward W. Gregg, PhD, Deborah Rolka, MS, and Meda E. Pavkov, MD PhD.

We thank Qilu Yu, Lori Merrill, and Sophia Jang for their assistance with data collection, harmonization, and programming support.

References

1. American Diabetes Association. Economic costs of diabetes in the U.S. in 2017. Diabetes Care. 2018;41(5):917–928. 10.2337/dci18-0007. [PubMed: 29567642]
2. Centers for Disease Control and Prevention. National diabetes statistics report: estimates of diabetes and its burden in the United States, 2017. Atlanta, GA: Centers for disease Control and prevention, US department of health and human services. https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf2017. Updated March 6, 2018. Accessed September 1, 2018.
3. Tabak AG, Herder C, Rathmann W, Brunner EJ, Kivimaki M. Prediabetes: a high-risk state for developing diabetes. Lancet. 2012;379(9833):2279–2290. [PubMed: 22683128]
4. Eggleston EM, Klompas M. Rational use of electronic health records for diabetes population management. Curr Diabetes Rep. 2014;14(4):479.
5. Anderson AE, Kerr WT, Thames A, Li T, Xiao J, Cohen MS. Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population: a cross-sectional, unselected, retrospective study. J Biomed Inf. 2016;60:162–168.
6. Nichols GA, Desai J, Elston Lafata J, et al. Construction of a multisite DataLink using electronic health records for the identification, surveillance, prevention, and management of diabetes mellitus: the SUPREME-DM project. Prev Chronic Dis. 2012; 9:E110. [PubMed: 22677160]
7. Health Resources & Services Administration (Hrsa). Federally qualified health center look-alike. Available at: https://www.hrsa.gov/opa/eligibility-and-registration/health-centers/fqhc-look-alikes/index.html; 2018.
8. Observational Health Data Sciences and Informatics (Ohdsi). OMOP Common Data Model. 2017.
9. 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;50:S21–S29. 10.1097/MLR.0b013e318257dd67. [PubMed: 22692254]
10. Mashayekhi M, Prescod F, Shah B, Dong L, Keshavjee K, Guergachi A. Evaluatingthe performance of the framingham diabetes risk scoring model in Canadian electronic medical records. Can J Diabetes. 2015;39:152–156. [PubMed: 25577729]
11. Chung S, Zhao B, Lauderdale D, Linde R, Stafford R, Palaniappan L. Initiation of treatment for incident diabetes: evidence from the electronic health records in an ambulatory care setting. Prim Care Diabetes. 2015;9(1):23–30. [PubMed: 24810147]
12. Kudyakov R, Bowen J, Ewen E, et al. Electronic health record use to classify patients with newly diagnosed versus preexisting type 2 diabetes: infrastructure for comparative effectiveness research and population health management. Popul Health Manag. 2012;15:3–11. [PubMed: 21877923]
13. Thomsen RW, Sorensen HT. Using registries to identify type 2 diabetes patients. Clin Epidemiol. 2014;7:>1–3. [PubMed: 25565888]
14. Eastwood SV, Mathur R, Atkinson M, Brophy S, Chaturvedi N. Algorithms for the capture and adjudication of prevalent and incident diabetes in UK biobank. PloS One. 2016;11(9). e0162388. [PubMed: 27631769]
15. Klompas M, Eggleston E, McVetta J, Lazarus R, Li L, Platt R. Automated detection and classification of type 1 versus type 2 diabetes using electronic health record data. Diabetes Care. 2013;36(4):914–921. [PubMed: 23193215]
16. Kirkman MS, Nooney JG, Benoit SR, et al. 211-LB: Developing" Gold Standard" Diagnoses for Type 1 and Type 2 Diabetes in Adults from Electronic Health Record Data. 2019.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

17. American Diabetes Association. Standards of medical care in diabetes—2016. Diabetes Care. 2016;39(suppl 1):S1–S112. [PubMed: 26696671]
18. Hostalek U, Gwilt M, Hildemann S. Therapeutic use of metformin in prediabetes and diabetes prevention. Drugs. 2015;75(10):1071–1094. [PubMed: 26059289]
19. Farr OM, Mantzoros CS. Treating prediabetes in the obese: are GLP-1 analogues the answer? Lancet. 2017;389(10077):371–372, 1.
20. Lashen H. Role of metformin in the management of polycystic ovary syndrome. Ther Adv Endocrinol Metab. 2010;1 (3): 117–128. [PubMed: 23148156]
21. Patel P, Macerollo A. Diabetes mellitus: diagnosis and screening. Am Fam Physician. 2010;81(7):863–870. [PubMed: 20353144]
22. Bansal N. Prediabetes diagnosis and treatment: a review. World J Diabetes. 2015;6(2): 296–303. [PubMed: 25789110]
23. Charfen MA, Ipp E, Kaji AH, Saleh T, Qazi MF, Lewis RJ. Detection of undiagnosed diabetes and prediabetic states in high-risk emergency department patients. Acad Emerg Med. 2009;16(5):394–402. [PubMed: 19302369]
24. Tentolouris N, Lathouris P, Lontou S, Tzemos K, Maynard J. Screening for HbA1c-defined prediabetes and diabetes in an at-risk Greek population: performance comparison of random capillary glucose, the ADA diabetes risk test and skin fluorescence spectroscopy. Diabetes Res Clin Pract. 2013;100(1):39–45. [PubMed: 23369230]
25. Jackson SL, Safo SE, Staimez LR, et al. Glucose challenge test screening for prediabetes and early diabetes. Diabet Med. 2016;34(5):716–724. [PubMed: 27727467]
26. SAS Institute, Inc, Version 9.4.
27. Tu K, Manuel D, Lam K, Kavanagh D, Mitiku TF, Guo H. Diabetics can be identified in an electronic medical record using laboratory tests and prescriptions. J Clin Epidemiol. 2011;64(4):431–435. [PubMed: 20638237]
28. Hersh W. Electronic health records facilitate development of disease registries and more. Clin J Am Soc Nephrol. 2011;6(1):5–6. [PubMed: 21127135]
29. Kadhim-Saleh A, Green M, Williamson T, Hunter D, Birtwhistle R. Validation of the diagnostic algorithms for 5 chronic conditions in the Canadian Primary Care Sentinel Surveillance Network (CPCSSN): a Kingston practice-based research network (PBRN) report. J Am Board Fam Med. 2013;26(2):159–167. [PubMed: 23471929]
30. Coleman N, Halas G, Peeler W, Casaclang N, Williamson T, Katz A. From patient care to research: a validation study examining the factors contributing to data quality in a primary care electronic medical record database. BMC Fam Pract. 2015;16:11. [PubMed: 25649201]
31. Antman EM, Benjamin EJ, Harrington RA, et al. Acquisition, analysis, and sharing of data in 2015 and beyond: a survey of the landscape—a conference report from the American Heart Association Data Summit 2015. J Am Heart Assoc. 2015;4(11). pii: e002810. [PubMed: 26541391]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

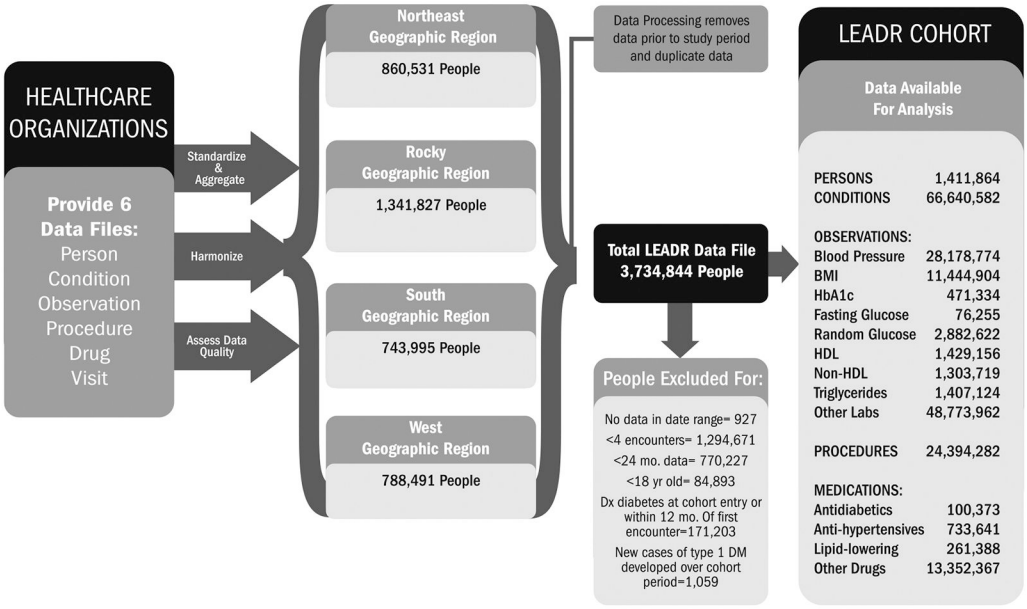

Fig. 1.
Building the LEADR cohort.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Healthc (Amst). Author manuscript; available in PMC 2024 April 11.

Table 1

Definitions.

Incident cases of T2DM	1. A diabetes diagnosis of T2DM or unspecified diabetes diagnosis made on two different days within 24 months (diagnosis dates ≥14 days apart) 2. A prescription for metformin or glucagon-like peptide-1 (GLP1) agonists and a T2DM or unspecified diabetes diagnosis on any encounter (activities < 14 days apart). 3. A prescription for an antidiabetic agent a) alpha-glucosidase inhibitors, b) amylin analogs, c)anti-diabetic agent combinations including those with metformin; d)insulin among non-pregnant women,6 e) meglitinides, f) sodium glucose cotransporter 2 (SGLT2) inhibitors, g) sulfonylureas, h) thiazolidinediones, and/or i) Dipeptidyl peptidase-4 (DDP-4) inhibitors 4) Lab results a) hemoglobin A1c lab result ≥6.5% or b) fasting plasma glucose ≥126 mg/dl, or c) i) random glucose ≥200 mg/dl on two different days within 24 months (diagnosis dates ≥ 14 days apart), or ii) random glucose ≥ 250 mg/dl, or iii) random glucose ≥ 200 mg/dl and a T2DM or unspecified diabetes diagnosis on any encounter (activities < 14 days apart).
Obesity	1. An obesity diagnosis 2. At least one body mass index (BMI) ≥ 30
BMI	Underweight, <18.5 kg/m2 Normal, 18.5 ≤ BMI < 25 kg/m2 Overweight, 25 ≥ BMI < 30 kg/m2 Obesity I, 30 ≥ BMI < 35 Obesity II, 35 ≤ BMI < 40 kg/m2 Obesity III, ≥ 40 kg/m2
Hypertension/ High Blood Pressure	1. Two hypertension diagnoses (≥14 days apart) 2. A hypertension diagnosis and a hypertension medication prescription a) angiotensin-converting enzyme inhibitors (ACE), b) angiotensin II receptor blockers (ARB), c) beta blockers, d) calcium channel blocks, and/or e) diuretics 3. A hypertension diagnosis and a) systolic blood pressure average ≥ 140 (if at least two results ≥ 14 days apart), or b) diastolic blood pressure average ≥ 90 (if at least two results ≥ 14 days apart)
Prediabetes/ Elevated Glycemia	1. A prediabetes diagnosis 2. Lab results a) hemoglobin A1c lab result ≥5.7% and <6.5% or b) fasting plasma glucose ≥100 mg/dL and <126 mg/dL
Elevated Lipids	1. An elevated lipids diagnosis 2. A prescription for elevated lipids medication a) statins or statin combinations b) fibrates c) niacin d) bile acid sequestrates, and/or e) other lipid-modifying agents

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

	3. Lab results a) triglyceride level ≥250 mg/dL b) HDL <40 mg/dL for males and <50 mg/dL for females. non-HDL value ≥160 mg/dL
Depression	A diagnosis of depression, including: a) major depressive disorder, or b) depression (mild, moderate, severe, endogenous, recurrent, single episode), or c) chronic depressive personality disorder, or d) major depressive affective disorder, or e) cyclothymic disorder, or f) dysthymia, or g) mixed anxiety and depressive disorder
History of Tobacco Use	1) A diagnosis of: a) tobacco dependence syndrome, or b) tobacco abuse, or c) tobacco user, or d) personal history of tobacco use, or e) nicotine dependence, or 2) Text denoting current tobacco user or quit smoking in Observation table, or 3) Treatment for tobacco use in Procedure table, or 4) A prescription for nicotine or varenicline
Family History of Diabetes	1. A diagnosis of family history of diabetes, or 2. A record in the Observation table denoting family history of diabetes

BMI = body mass index; HDL = high density lipoprotein; T2DM = type 2 diabetes

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Healthc (Amst). Author manuscript; available in PMC 2024 April 11.

Table 2

Data Elements Collected for 1.4 million LEADR Cohort Patients, 2010–2016.

Record Type	N (records)	% of patients with 0 records	N (patients with ≥1 record)	Mean # of records/ patient[a]	Percentile[a]		
					5th	50th	95th
Condition/Diagnosis	66,640,582	2%	1,386,116	48	3	25	168
Body Measurements							
Blood pressure[b]	28,178,774	17%	1,166,407[c]	12	1	7	39
BMI	11,444,904	35%	922,179	12	1	8	39
Laboratory Results							
HbA1c	471,334	83%	236,320	2	1	1	5
Fasting plasma glucose	76,255	98%	32,332	2	1	1	8
Random glucose	2,882,622	59%	584,526	5	1	3	15
High-density lipoprotein	1,429,156	65%	489,356	3	1	2	8
Non-high-density lipoprotein	1,303,719	66%	474,663	3	1	2	8
Triglycerides	1,407,124	64%	503,075	3	1	2	8
Other labs[d]	48,773,962	28%	1,013,770	48	1	14	239
Medications							
Drug - special[e]	1,095,402	64%	508,994	2	1	2	5
Drug - other[f]	13,352,367	8%	1,300,717	10	1	7	30
Procedures[g]	24,394,282	25%	1,062,867	23	1	11	85
Referrals							
Nutrition/Exercise Counseling	275,413	93%	106,466	3	1	1	7
Health Behavior Intervention/Risk Factor Reduction	190,792	95%	68,186	3	1	1	10

BMI = body mass index; HbA1c = hemoglobin A1c.

[a]Mean and percentiles are calculated for patients with at least one record of the specified type.

[b]There is not complete correspondence between systolic and diastolic blood pressure measurements. However, the statistics displayed here match for both systolic and diastolic measurements. Of the 28,178,774 blood pressure records, 14,148,706 (50.2%) are diastolic and 14,030,068 (49.8%) are systolic.

[c]There are fewer patients with at least one systolic blood pressure measurement, 1,165,562.

[d]Other labs include laboratory measurements of blood, serum, and urine, including blood cell counts, blood chemistry, and markers of liver and kidney function such as alanine aminotransferase, aspartate aminotransferase, and estimated glomerular filtration rate, and numerous other measurements.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[e] Medications of special interest include antidiabetics, lipid lowering medications, and antihypertensive medications.

[f] The top fifteen other medications are ibuprofen 800 mg, omeprazole 40 mg, fluticasone propionate 0.05 mg, azithromycin 250 mg, ibuprofen 600 mg, zolpidem tartrate 5 mg, omeprazole 20 mg, naproxen 500 mg, gabapentin, prednisone 20 mg, oxycodone hcl, levothyroxine sodium, ondansetron 4 mg, aspirin 500 mg, acetaminophen 500 mg/hydrocodone bitartrate 5 mg. Each represents <1% of all records.

[g] Procedures are medical procedures such as ultrasounds and colonoscopies.

Healthc (Amst). Author manuscript; available in PMC 2024 April 11.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3

LEADR cohort demographics, regional distribution, and incident cases of diabetes.

Characteristic	N	%
Total	1,411,864	100.0
Race/Ethnicity		
Non-Hispanic White	989,750	70.1
Non-Hispanic Black	119,989	8.5
Hispanic	135,401	9.6
Other/Multi-racial	11,928	0.8
American Indian/Alaska Native	5210	0.4
Asian	22,541	1.6
Missing/Unknown/Refused	127,045	9.0
Age, years (as of 2010)		
18 to 24	145,279	10.3
25 to 34	232,562	16.5
35 to 44	240,909	17.1
45 to 54	292,573	20.7
55 to 64	254,779	18.1
65 and older	245,762	17.4
Sex		
Female	855,937	60.6
Male	555,896	39.4
Missing/Unknown/Refused	31	<0.1
Geographic Region		
Northeast	486,585	34.5
Rocky	499,764	35.4
South	249,319	17.7
West	176,196	12.5
Type 2 Diabetes Mellitus Incident Cases (developed over study period)		
No	1,328,942	94.1
Yes	82,922	5.9