Development of Portable Electronic Health Record Based Algorithms to Identify Individuals with Diabetic Retinopathy

Joseph H. Breeyear[1,2,3], Sabrina L. Mitchell[3,4], Cari L. Nealon[5], Jacklyn N. Hellwege[1,3,6], Brian Charest[29], Anjali Khakharia[7,8], Christopher W. Halladay[9], Janine Yang[10], Gustavo A. Garriga[11], Otis D. Wilson[3,28], Adriana M. Hung[3,28], Peter D. Reaven[12,13], James B Meigs[14,15,16], Mary K Rhee[7,17], Yang Sun[18], Mary G. Lynch[7], Lucia Sobrin[10], Milam A. Brantley, Jr.[1,3,4], Yan V. Sun[7,20,21], Peter W. Wilson[7,22], Sudha K. Iyengar[23,24], Neal S. Peachey[23,25,26], Lawrence S. Phillips[7,17], Todd L. Edwards[1,3,27], Ayush Giri[1,3,11,27]

[1]Vanderbilt Genetics Institute, Vanderbilt University, Nashville, TN
[2]Biostatistics and Computational Biology Branch Division of Intramural Research, National Institute of Environmental Health Sciences, Durham, NC
[3]VA Tennessee Valley Healthcare System (626), Nashville, TN
[4]Department of Ophthalmology and Visual Sciences, Vanderbilt University Medical Center, Nashville, TN
[5]Eye Clinic, VA Northeast Ohio Healthcare System, Cleveland, OH
[6]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN
[7]VA Atlanta Healthcare System, Decatur, GA
[8]Department of Medicine and Geriatrics, Emory University School of Medicine, Atlanta, GA
[9]Providence VA Medical Center, Providence, RI
[10]Department of Ophthalmology, Mass Eye and Ear Infirmary, Harvard Medical School, Boston, MA
[11]Division of Quantitative Sciences, Department of Obstetrics and Gynecology, Vanderbilt University Medical Center, Nashville, TN
[12]Phoenix VA Health Care System, Phoenix, AZ
[13]College of Medicine, University of Arizona, Phoenix, AZ
[14]Program in Medical and Population Genetics, Broad Institute, Cambridge, MA
[15]Department of Medicine, Harvard Medical School, Boston, MA
[16]Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA
[17]Division of Endocrinology, Metabolism, and Lipids, Department of Medicine, Emory University School of Medicine, Atlanta, GA
[18]Department of Ophthalmology, Stanford University School of Medicine, Palo Alto, CA
[19]Department of Epidemiology, Emory University Rollins School of Public Health, Atlanta, GA
[20]Department of Epidemiology, Emory University Rollins School of Public Health, Atlanta, GA
[21]Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA
[22]Division of Cardiology, Department of Medicine, Emory University School of Medicine, Atlanta, GA
[23]Research Service, VA Northeast Ohio Healthcare System, Cleveland, OH
[24]Department of Population and Quantitative Health Sciences, Case Western Reserve University School of Medicine, Cleveland, OH
[25]Cole Eye Institute, Cleveland Clinic, Cleveland, OH
[26]Department of Ophthalmology, Cleveland Clinic Lerner College of Medicine of Case Western Reserve University, Cleveland, OH
[27]Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN
[28]Division of Nephrology and Hypertension, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN
[29]Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC), VA Boston Healthcare System, Boston, MA

Address for Co-Correspondence:

Ayush Giri

2525 West End Avenue
6th Floor, Suite 600, Office 631
Nashville, TN, 37203

Todd L. Edwards

2525 West End Avenue
6th Floor, Suite 600, Office 615
Nashville, TN, 37203

Telephone: (615) - 343 – 4553, (615) 322 – 3652
Email: ayush.giri@vumc.org, todd.l.edwards@vumc.org

Word Count: 4,050

ABSTRACT

Objectives/Aims: To develop, validate and implement algorithms to identify diabetic retinopathy (DR) cases and controls from electronic health care records (EHR)s. **Methods**: We developed and validated EHR-based algorithms to identify DR cases and individuals with type I or II diabetes without DR (controls) in three independent EHR systems: Vanderbilt University Medical Center Synthetic Derivative (VUMC), the VA Northeast Ohio Healthcare System (VANEOHS), and Massachusetts General Brigham (MGB). Cases were required to meet one of three criteria: 1) two or more dates with any DR ICD-9/10 code documented in the EHR, or 2) at least one affirmative health-factor or EPIC code for DR along with an ICD9/10 code for DR on a different day, or 3) at least one ICD-9/10 code for any DR occurring within 24 hours of an ophthalmology exam. Criteria for controls included affirmative evidence for diabetes as well as an ophthalmology exam. **Results**: The algorithms, developed and evaluated in VUMC through manual chart review, resulted in a positive predictive value (PPV) of 0.93 for cases and negative predictive value (NPV) of 0.97 for controls. Implementation of algorithms yielded metrics that were similar in VANEOHS (PPV=0.94; NPV=0.86) and lower in MGB (PPV=0.84; NPV=0.76). In comparison, the algorithm for DR implemented in Phenome-wide association study (PheWAS) in VUMC, yielded similar PPV (0.92) but substantially reduced NPV (0.48). Implementation of the algorithms to the Million Veteran Program identified over 62,000 DR cases with genetic data including 14,549 cases who identify as African American and 6,209 cases who identify as Hispanics. **Conclusions/Interpretations**: We demonstrate the robustness of the algorithms at three separate health-care centers, with all sites achieving a PPV of 0.84 or greater and substantially improved NPV than existing automated methods. We show EHR algorithms for DR are portable between separate EHR systems and also discuss lessons learned from these implementations. Algorithms reported here are advanced starting points. We strongly encourage independent validation and incorporation of features unique to each EHR to enhance algorithm performance for DR cases and controls.

KEYWORDS: Diabetic Complications, Diabetic Retinopathy, Algorithm Development, Algorithm Validation

ABBREVIATIONS: Diabetic Retinopathy [DR], Proliferative Diabetic Retinopathy [PDR], Electronic Health Record [EHR], Diabetes Mellitus [DM], Million Veteran Program [MVP], Positive Predictive Value [PPV], Negative Predictive Value [NPV]

RESEARCH IN CONTEXT (Bullet point word count: 200 words)

What is already known about this subject?

- Genetic investigations of diabetic retinopathy [DR] have been small due to difficulty of identifying cases and controls.
- Use of electronic health care records (EHRs] to complement research endeavors for diabetic complications are increasing, with vigilant skepticism.
- The utility of diagnostic/procedural code-based algorithms routinely collected in EHRs in identifying DR cases and controls remains unclear.

What is the key question?

- Can we use diagnostic and procedural code-based algorithms to accurately identify DR cases and controls in multiple EHR systems?

What are the new findings?

- Algorithms developed at Vanderbilt University Medical Center yielded high predictive values upon manual chart review (positive predictive value [PPV] for DR: 0.94; negative predictive value [NPV] for diabetic controls: 0.97); its performance exceeds the algorithm implemented in the popular Phenome wide association study methodology [PPV = 0.92; NPV = 0.48]
- Application of similar algorithms yielded high, yet variable, accuracies at the VA Northeast Ohio Healthcare System (VANEOHS) [PPV = 0.94; NPV = 0.86], and Massachusetts General Brigham (MGB) [PPV = 0.84; NPV = 0.76]
- Application to the Million Veteran Program allowed identification of over 62,000 DR cases and 110,000 controls, which will enable large-scale investigations of DR etiology.

How might this impact on clinical practice in the foreseeable future?

- Thoughtful application of algorithms to EHRs offers the opportunity to ask previously intractable research/clinical questions pertaining to DR etiology.

INTRODUCTION

Diabetic retinopathy (DR), which can manifest with breakdown of the blood-retinal barrier and neovascularization in the retina, is a leading cause of visual impairment and preventable blindness among working-age adults, affecting an estimated 103 million individuals worldwide.[1] As the global prevalence of diabetes mellitus (DM) rises, the global prevalence of DR will also rise, with estimates suggesting there will be more than 160 million individuals with DR by 2045.[1]

An extensive network of intraretinal vasculature maintains intricate control of homeostasis in the ocular environment. High metabolic rate induced by DM makes these delicate structures highly susceptible to damage.[2] The prevalence of DR varies between individuals with type I and type II diabetes, but has been estimated to be between 37% and 94% for type I DM and 35% for type II DM.[3] Among type I and type II individuals, nearly one-third of those with DR progress to vision-threatening DR over years of disease, respectively.[2-4] Early-stage DR, termed non-proliferative DR, is characterized by blot hemorrhages, cotton wool spots, intraretinal microvascular abnormalities, retinal microaneurysms, and venous beading. Based on the extent of retinal findings, non-proliferative DR is classified as mild, moderate, or severe. The more advanced proliferative diabetic retinopathy (PDR) is defined by neovascularization of the retina or iris. Other serious ocular conditions can arise during progression of DR including diabetic macular edema, which can occur with either non-proliferative DR or PDR, and neovascular glaucoma which occurs secondary to PDR.

DR is a complex disease with multiple genetic and environmental factors contributing to the pathophysiology. The strongest established risk factors for DR and for progression to vision-threatening DR are longer DM duration and poor glycemic control, while other reported risk factors include hypertension, diabetic neuropathy, diabetic nephropathy, and increased body mass index [BMI].[2, 5, 6] Compared with populations of European descent, the prevalence of DR and vision-threatening DR is higher in African American and Hispanic populations.[7-11] Conducting epidemiologic and genetic studies of DR based on manual review of retinal imaging, manual review of electronic health record (EHR) data, or prospective cohorts has been

challenging, leading to smaller sample sizes. Additionally, many prospective studies have been performed primarily in individuals of European ancestry.

At present, only a small number of genetic loci have been associated with DR.[12-19] These have been identified in studies with recruitment following face-to-face encounters, where cases are individuals with diabetes and DR, while controls are individuals with diabetes without DR. Large-scale biobanks provide important opportunities to pursue additional loci and to identify genetic and lifestyle related risk factors for DR. To leverage these resources for understanding etiology behind DR, an efficient and accurate means to identify individuals with diabetes and DR as well as individuals with diabetes without DR is required. Automated approaches such as Phenome-wide association study (PheWAS) have provided the opportunity to leverage EHR systems for research, however, the validity of the algorithm utilized, especially in the context of DR, has not been evaluated to our knowledge. A validated DR algorithm that can be applied in EHRs would increase analytical sample sizes, thereby improving statistical power and enabling research groups to ask previously intractable research questions to increase our understanding of the underlying pathophysiology of this potentially severe DM complication. To date, a single validated EHR-based algorithm for DR has been presented in the literature, one that was developed and validated in a single EHR system.[20] Here, we describe the development and portability of EHR-based algorithms to identify individuals with DR (cases) and individuals with DM but without DR (controls) across three healthcare systems in the US.

RESEARCH DESIGN and METHODS

In this study, we included EHR data from three independent healthcare systems, the Vanderbilt University Medical Center (VUMC Synthetic Derivative (SD)), the VA Northeast Ohio Healthcare System (VANEOHS) Eye Clinic, and the Massachusetts General Brigham (MGB) biobank. This study was conducted upon approval of Institutional Review Boards of VUMC, VANEOHS, MGB, and Million Veteran Program (MVP) central and local offices.

Algorithm Development

The algorithm development was performed in the VUMC SD, a database containing de-identified clinical information derived from VUMC's electronic medical records, described elsewhere.[21] Briefly, the VUMC SD contains records for >2.2 million individuals and includes diagnostic and procedure codes, demographics (age, sex, race), clinical notes, laboratory values, and other medication information. The algorithm development was an iterative process, where in **Version 0** (**V0**) required that cases have at least one ICD-9/10 code for any form/severity of DR, while controls were required to have at least one ICD-9/10 code for DM and no codes from the DM exclusion list (**Supplemental Tables 1 – 2**) and at least one eye exam and had no codes for DR. In **V1**, we required cases to have at least two unique dates with an ICD-9/10 code for any form of DR, and controls to have at least three dates with an ICD-9/10 code for DM without any DR codes. This additional requirement for control was put in place to improve confidence that a given patient had diabetes during chart review. In **V2**, we relaxed the algorithm by adding an eligibility criterion allowing cases to have a single ICD-9/10 code for any form of DR if accompanied within 24 hours of an ophthalmology exam. Along with the DR algorithm, we developed an algorithm for PDR based on restricting criteria in the "any DR" algorithm to only PDR ICD-9/10 codes.

Algorithm Portability

VA Northeast Ohio Healthcare System

In the VANEOHS EHR, individuals with DM were first identified using a combination of diagnosis and procedure codes, clinic stop codes, and medications. Briefly, individuals with affirmative evidence of DM were identified based on presence of a diagnosis code at one or more face-to-face outpatient primary care physician visits or a DM diagnosis code at two or more outpatient visits of any kind, and at least one outpatient fill of a diabetes-indicated medication (**Supplemental Table 3**). In addition to ICD9/10 codes the VA also routinely utilizes Health Factor codes to annotate health records. To take advantage of this additional information source, in addition to the DR case definition (**V2**), we allowed VA cases to have a Health Factor code along with an ICD-9/10 code for any form of DR on unique days. The algorithm was subsequently applied in the U.S. Department of Veterans Affairs biorepository, the Million Veteran Program.

Mass General Brigham

In the MGB, individuals with DM were first identified with a previously curated DM phenotype.[22] To identify DR cases, we applied the DR case definition (**V2**) with the additional criteria of an EPIC Provider code along with an ICD-9/10 code for any form of DR on unique days, to take advantage of features unique to MGB. We used the strict algorithm for controls in MGB (**V1**), which requires at least three unique days of diabetes codes to be included as controls, as applying only one ICD code for diabetes (**V0**) yielded a large proportion of records where diabetes status could not be validated with manual review.

Algorithm Evaluation

In the VUMC SD, we conducted a blinded manual review of 495 de-identified EHRs by an ophthalmologist and 4 other trained reviewers, identifying DR status, DR severity, and control status, and compared the results to the algorithm definitions (**Supplemental Figure 1**). Of the 495 records manually reviewed, 45 de-identified EHRs were overlapping to assess concordance among the reviewers. Performance of the algorithms was primarily assessed by computing positive predictive values (PPV) for algorithmically determined cases, negative predictive values (NPV) for algorithmically determined controls. Sensitivity and specificity were computed from resulting tabulations. If an algorithmically determined case or control could not be validated with manual review, they were classified as undetermined and were excluded from the numerator but included in the denominator for computing NPV. For the 45 overlapped records, we only utilized one manual review. If the reviewers disagreed, we used a conservative approach to include the manual review that did not bias our results towards a higher PPV or NPV. As we could not manually review retinal imaging data for individuals in the VUMC SD, we captured a subset of EHRs (n = 152) from the 495 that included EHR notes from the Vanderbilt Eye Institute (VEI). These VEI notes indicate that an individual was assessed in the clinic with a retinal exam with an ophthalmologist reporting on their findings. Of the 45 overlapped EHRs, 21 had VEI notes. Finally, in the VUMC SD, we compared the performance of our algorithm with the widely used PheCode definition of DR, by evaluating test characteristics DR cases and controls identified using the PheCode definition.[23]

At the VANEOHS, we pulled 50 algorithm-defined cases and 50 algorithm-defined controls for chart review. Charts were reviewed by a single optometrist that was blinded to the algorithm assignments. Twenty-one additional charts were pulled for further review to improve precision of the PDR algorithm since there were too few PDR cases in the first 50 records with any-DR. Of the 50 any-DR cases, we identified 32 visits for 17 DR patients for whom at least one fundus imaging was available for review. We investigated the performance of the algorithm against these fundus images manually reviewed by an optometrist.

In the MGB biobank, we pulled 50 algorithm-defined cases (**V2**) and 50 algorithm-defined controls (**V1**) for chart review. Charts were reviewed by a fellowship-trained retinal specialist that was blinded to the algorithm assignments.

Phenome-Wide Association Study

We performed a phenome-wide association study (PheWAS) of the DR algorithm (**V2**) in VUMC SD (n_{max} = 12,991) and the MVP (n_{max} = 165,322) leveraging the full catalog of ICD-9/10 codes.[24] The ~1,800 phecode traits were modeled with logistic regressions as a function of the algorithm (**V2**) designation, adjusted for age, sex, and the duration of DM, stratified by self-reported race/ethnicity (non-Hispanic White (NHW), non-Hispanic Black (NHB) , and Hispanic (HIS)). Interpretation of results were limited to phenotypes with 20 or more cases. Bonferroni-corrected significance threshold was set to p - value $\leq 2.75 \times 10^{-5}$ to account for the total number of PheCodes assessed. The results were meta-analyzed both within and across populations. Analyses were conducted in R with the PheWAS package.[25]

RESULTS

We compared the performance of the three DR algorithm versions developed in the VUMC SD dataset (**Table 1**). The initial algorithm (**V0**), aimed at identifying possible cases and controls, resulted in a PPV and NPV of 0.857 and 0.934, respectively. The second algorithm (**V1**), designed to increase algorithm precision, resulted in a substantially improved PPV (0.969) and provided minimal gains on NPV (0.966) compared to the V0 versions. The final algorithm for cases (**V2**), intended to maximize cases and maintain a high PPV by including ophthalmology exam data along with a single DR ICD code instead of two, resulted in a similar PPV

(0.930) as **V1**, while increasing case-numbers. Compared to the PheCode definition (PPV = 0.922, NPV = 0.480), our algorithm classified cases (**V2**) with a similar PPV, but a considerably higher NPV (**V0** and **V1**). The same upward trend was seen for the PDR algorithm, with the PPV increasing from 0.802 to 0.914 (**Table 1**). In the subset of EHRs with VEI notes, we saw similar algorithm performance for PPV (0.938) and improved performance for NPV (1.0). The concordance of the five reviewers for DM, DR, and PDR designations was 96%, 91%, and 97%, respectively (**Supplemental Table 4**). When restricted to EHRs with notes from the Vanderbilt Eye Institute, the concordance for DM, DR, and PDR designations was 100%, 95%, and 100% of the individuals, respectively (**Supplemental Table 4**).

The algorithms performed well in VANEOHS, where the DR algorithm PPV was 0.940 and PDR algorithm PPV was 0.818 (**Table 1, Supplemental Table 5**). There was a slight drop in the NPV for the VANOEHS (NPV = 0.860), while the NPV for the PDR algorithm (NPV = 0.989) was comparable to the VUMC. Among the 50 algorithmically-defined any-DR cases, we found fundus photography images associated with 32 visits for 17 patients. On a visit-level, 29 of 32 (PPV = 0.90) algorithmically-defined cases were verified to have DR by manual review of imaging. Three DR-determinations could not be verified directly via imaging due to poor image quality. Of the 29 visits with adequate image quality, staging of DR (background/mild non-proliferative diabetic retinopathy, moderate non-proliferative diabetic retinopathy, severe non-proliferative diabetic retinopathy, and proliferative diabetic retinopathy) was in agreement for 25 images between our manual review and algorithm-determined diagnosis (Supplemental Table 10). The algorithm tended to underestimate the staging of DR compared to manual review for the other remaining four images. On a patient-level 16 of the 17 patients (PPV = 0.94) were verified as DR cases through manual review of images. Application of the algorithm to MGB showed lower PPV (0.840) for **V2** cases and lower NPV (0.760) for **V1** controls compared to VUMC SD and VANEOHS.

We applied these algorithms to identify DR cases and DM controls in three separate EHR systems; 8,860 DR cases (**V2**) and 18,654 controls (**V0**) in the VUMC SD, 62,101 DR cases (**V2**) and 110,688 controls (**V0**)

in the MVP, and 471 DR cases (**V2**) and 1,137 controls (**V1**) in the MGB (**Table 2**). The three cohorts differed substantially in their sex composition with VUMC SD and MGB having approximately 50% females, whereas MVP consisted of predominantly males (>95%) (**Table 2**). Despite these drastic differences in cohort compositions, seminal observations for DR cases and controls were consistent with those in epidemiologic literature, where individuals with DR on average were more likely to have diabetic macular edema, neovascular glaucoma, diabetic neuropathy, and diabetic nephropathy in both the VUMC SD and MVP EHR systems. By design, diabetics without DR were more likely to be older than DR cases across all three EHR systems. Average duration of diabetes, captured as time since first diagnosis of diabetes in EHR or first mention of antihypertensive medication in EHR, whichever came first, was similar between VUMC and MVP cases, while controls in the MVP had a longer duration than in VUMC. At VUMC SD, controls (V1) were more likely to have diagnostic codes for type II diabetes, more likely to have hypertension than DR cases and more likely to have ophthalmic visits in the last five years. Such phenomenon was not observed in the MVP EHR system. Cases and controls were similar on average with regard to BMI and had similar proportions of inherited retinal disease in VUMC SD and MVP.

Phenome-Wide Association Study

To systematically evaluate the relationship of algorithmically-defined DR cases and controls with the phenome derived from EHR, we performed a Phenome-wide association study across study populations in the MVP and SD. The DR algorithm PheWAS identified a total of 959 unique PheCodes significantly associated with DR, with 99, 8, 8, and 7 significantly associated in only the Multi, NHW, NHB, and HIS populations, respectively (**Table 3, Supplemental Tables 6 - 9**). In the Multi-ancestry meta-analysis, a total of 784 PheCodes were associated with increased DR risk while 151 were associated with decreased DR risk. The most significantly associated PheCodes included Diabetic Retinopathy, Type 2 DM with Ophthalmic Manifestations, and Type 1 DM with Ophthalmic Manifestations, all with p – values below 1.00×10^{-305}. In addition, the DR algorithm designation was associated with other phenotype groups related to downstream consequences of DM in the

multi-ancestry meta-analysis, including metabolic disorders (82), cardiovascular disorders (127), and renal disorders (70).

DISCUSSION

In this study, we repurposed diagnostic codes and administrative data routinely collected in clinical practice to develop algorithms that identify individuals with diabetes, with and without diabetic retinopathy, to be used for research purposes. We involved three separate healthcare systems to confirm the portability of the DR algorithms, along with a comparison of the algorithm with manually reviewed fundus imaging. While identification of individuals with DR typically relies on manual review or machine learning algorithms along with retinal imaging data, our algorithm focused on ICD diagnostic codes and is thus transferable to other sites, allows for modifications to capitalize on unique EHR features, and requires significantly fewer resources for identifying at-risk individuals. The DR algorithms tested accurately identified individuals with DR, with PPVs ranging from 0.840 – 0.944 and identified those with diabetes and without DR with NPVs ranging from 0.76 to 0.97. This is a vast improvement over existing algorithms implemented in high-throughput methods such as PheWAS (PPV = 0.93; NPV = 0.48), and represents a starting point for researchers looking to maximize use of EHR for DR research.

To date, most studies have used one of three methods to identify individuals with DR: 1) manual review of eye examination findings, (2) manual review of retinal imaging, or (3) machine learning algorithms to review retinal imaging data.[12, 17, 26, 27] While these methods are considered the gold standard for phenotyping, they are labor intensive and can require retinal imaging data, which is not always available in research datasets. The largest DR genome-wide association study to date, including 19,372 cases and 19,615 controls, involved the manual review of retinal images for more than 38,987 individuals from multiple studies.[12] As studies of DR are often under-powered, a method to easily identify individuals with DR with relatively minimal effort while maintaining high accuracy through EHRs offers an attractive opportunity to scale up research related to DR. We first show through manual review of records that the algorithms accurately identify individuals coded as DR cases and controls. Following the premise that DR diagnoses of

patients seen by eye-specialists would be most reliable, we stratified cases and controls in the VUMC SD chart review validation by whether they visited the Vanderbilt Eye Institute. We showed patients who had VEI notes had higher, yet comparable, accuracy as cases and controls who were not seen by eye-specialists suggesting the high PPVs and NPVs were not entirely driven by one stratum. Then in the VA, we verified high PPV of our any-DR case-algorithm by manually reviewing 32 fundus images for 17 of the 50 any-DR cases for whom images were stored in the VA system. Further evaluation of DR-staging between algorithm and manual review showed high accuracy with close matches for 25 of 29 images with good quality. Reasons for incomplete availability of imaging include: the VA does not systematically store fundus photography imaging; imaging could have been done at a different VA location and the images would not be available at the local VA where the review was conducted, or the diagnosis may have come from outside of the VA system.

As with any study incorporating EHR data, there is a possibility of incomplete data, since individuals may receive care outside of the hospital system. Our experiences of developing and validating algorithms for DR cases and controls across three heterogeneous EHR systems has led to considerations and insights worth reflecting in the usage and interpretation of data derived from these sources, especially in the context of DR (catalogued in **Table 4**). Electronic health record systems, designed for healthcare utilization, often have center-specific methods and customizations, even within those using a single EHR platform (i.e. EPIC). As was the case in our study, data accessibility across EHR systems vary greatly either due to heterogenous data collection systems, provider practices, or due to data availability issues imposed across each EHR system. For example, in the MGB we found a sizable proportion of individuals were missing adequate EHR notes to validate ICD diagnoses. At the expense of losing controls, we limited evaluation of records with 3 or more diabetes diagnosis codes (**V1** control algorithm). While this alleviated the problem and still, 13% of individuals were missing adequate EHR notes needed for validation. Since these undetermined individuals were conservatively counted against the PPV and NPV to avoid inflation of predictive values, the resulting PPV and NPV are noticeably lower than the VUMC and VANEOHS.

Furthermore, we found that in addition to finding common features across systems, embracing differences between systems to retain as much detail as available greatly improves portability and validity of these algorithms. For example, in addition to the standard ICD9/10 diagnostic criteria, that formed the backbone for the primary algorithm at VUMC, the VANEOHS and the MGH systems each had additional coding systems that were unique to them, which we found more beneficial to incorporate rather than ignore. Limiting our algorithm to ICD9/10 codes and ignoring the Health Factor coding system in the VA would have led to more than 5,000 individuals misclassified as controls. Similarly, we were able to capitalize on the richness of the VA EHR, which included detailed information on prescription, and the type of prescribing physician (primary care provider, specialist, inpatient, outpatient, etc.) to gain affirmative evidence of diabetes status, features that were not accessible for the VUMC SD or MGH.

We tested features of the algorithms which have consequences on predictive properties, number of individuals identified and the composition of individuals identified. Requiring two or more unique days of diagnostic code for DR substantially improved PPV, but it also reduced the number of cases identified. Relaxing the criteria to one or more ICD and requiring evidence of eye-exam within 24 hours of the diagnostic code maintained previous gains in PPV from V1 and improved the number of cases. At VUMC, we found requiring at least three unique days of codes for diabetes for controls (compared to at least one), marginally improved NPV and substantially reduced sample size. On the other hand, in MGH we found requiring at least three unique days of diagnostic codes specifically for diabetes was necessary for optimal performance of control selection. EHRs with limited availability of information on prescription (inpatient vs. outpatient) and details regarding prescriber (primary care provider, specialist or otherwise) or chart notes, other than structured ICD codes may benefit with these criteria.

Comparative analyses across versions of algorithms within and across institutions also led to substantive insights. We observe a higher proportion of controls had T2D diagnostic codes in VUMC than DR cases, a phenomenon not observed in the VA. This is likely reflective of VUMC as a tertiary care facility, to which severe cases DR are referred for management. We note requiring at least three diabetes diagnostic

codes for controls vs. one, further changes characteristics of controls, as it prioritizes individuals who are older, have longer duration of diabetes, and are more likely to have other complications related to diabetes than controls with one ICD code. The VA's role as a primary care facility for veterans is further evidenced in our selection of patients where both cases and controls are just as likely to have T2D diagnosis codes and just as likely to have had eye exams in the past five years. Duration of diabetes is a particularly difficult variable to quantify through EHR alone in the US as it is only as accurate as the first recorded ICD code/diabetes prescription in each EHR system. This was evident in both VUMC and MVP where for several individuals the first mention of T2D diagnosis came after mention of a DR code, suggesting individuals were either only identified as diabetics after they developed complications or that their diagnosis of diabetes from external institutions were not captured. To ensure controls with diabetes did not have DR but had the opportunity of being diagnosed with DR, we imposed requirements for controls including ophthalmology/eye clinic and capturing age at last visit, whereas cases' ages were recorded at first mention of DR diagnosis. Consequently, we observe controls on average are older than cases across all three EHR systems. This strategy greatly reduced misclassification of outcome, which is highly desirable in the context of GWAS studies.

Overall, by developing and evaluating algorithms to identify diabetics with and without DR, we show the potential of leveraging existing EHR systems to advance our understanding of etiology of DR. We demonstrate these EHR-based algorithms to be portable and accurate across separate EHR systems and this represents a vast improvement over existing high-throughput method such as PheWAS. These algorithms provide advanced starting points for researchers looking to implement DR algorithms in EHR systems and rather than blind-automated application, we encourage researchers to investigate unique features in each EHR system to implement choices most appropriate to that system. As EHR systems increase in size, as well as the usage for research, a clear advantage to this approach is scalability as it offers an alternative to the manual review of large numbers of retinal images for research purposes while retaining validity.

ACKNOWLEDGEMENTS

This research is based on data from the Million Veteran Program, Office of Research and Development, and Veterans Health Administration. This publication does not represent the views of the Department of Veterans Affairs or the United States Government. We do not report any conflicts of interest. A.G. and T.L.E are the guarantors of this work and, as such, had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

FUNDING

Efforts were supported by the NIH/NEI F31EY033663 (J.H.B.), K12HD043483 (J.N.H.), R01-EY025295 (Y.S.), R01-EY032159 (Y.S.), 1K01DK120631 (A.G.), the VA Office of Research and Development [I01 CX001298 (Y.S.), I01CX001481 (Y.S.), I01BX005831 (L.S.P), I01BX004557 (N.S.P.), IK6BX005233 (N.S.P.)] and unrestricted awards from Research to Prevent Blindness.

REFERENCES

[1] Teo ZL, Tham YC, Yu M, et al. (2021) Global Prevalence of Diabetic Retinopathy and Projection of Burden through 2045: Systematic Review and Meta-analysis. Ophthalmology 128(11): 1580-1591. 10.1016/j.ophtha.2021.04.027

[2] Sabanayagam C, Banu R, Chee ML, et al. (2019) Incidence and progression of diabetic retinopathy: a systematic review. Lancet Diabetes Endocrinol 7(2): 140-149. 10.1016/s2213-8587(18)30128-1

[3] Lee R, Wong TY, Sabanayagam C (2015) Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss. Eye Vis (Lond) 2: 17. 10.1186/s40662-015-0026-2

[4] Zhang X, Saaddine JB, Chou C-F, et al. (2010) Prevalence of Diabetic Retinopathy in the United States, 2005-2008. JAMA 304(6): 649-656. 10.1001/jama.2010.1111

[5] Yau JW, Rogers SL, Kawasaki R, et al. (2012) Global prevalence and major risk factors of diabetic retinopathy. Diabetes Care 35(3): 556-564. 10.2337/dc11-1909

[6] Cheung N, Mitchell P, Wong TY (2010) Diabetic retinopathy. Lancet 376(9735): 124-136. 10.1016/s0140-6736(09)62124-3

[7] Lamoureux EL, Hassell JB, Keeffe JE (2004) The impact of diabetic retinopathy on participation in daily living. Arch Ophthalmol 122(1): 84-88. 10.1001/archopht.122.1.84

[8] Muñoz B, West SK, Rubin GS, et al. (2000) Causes of blindness and visual impairment in a population of older Americans: The Salisbury Eye Evaluation Study. Arch Ophthalmol 118(6): 819-825. 10.1001/archopht.118.6.819

[9] Coyne KS, Margolis MK, Kennedy-Martin T, et al. (2004) The impact of diabetic retinopathy: perspectives from patient focus groups. Fam Pract 21(4): 447-453. 10.1093/fampra/cmh417

[10] Varma R, Torres M, Peña F, Klein R, Azen SP (2004) Prevalence of diabetic retinopathy in adult Latinos: the Los Angeles Latino eye study. Ophthalmology 111(7): 1298-1306. 10.1016/j.ophtha.2004.03.002

[11] Muñoz B, O'Leary M, Fonseca-Becker F, et al. (2008) Knowledge of diabetic eye disease and vision care guidelines among Hispanic individuals in Baltimore with and without diabetes. Arch Ophthalmol 126(7): 968-974. 10.1001/archopht.126.7.968

[12] Pollack S, Igo RP, Jr., Jensen RA, et al. (2019) Multiethnic Genome-Wide Association Study of Diabetic Retinopathy Using Liability Threshold Modeling of Duration of Diabetes and Glycemic Control. Diabetes 68(2): 441-456. 10.2337/db18-0567

[13] Xue Z, Yuan J, Chen F, et al. (2022) Genome-wide association meta-analysis of 88,250 individuals highlights pleiotropic mechanisms of five ocular diseases in UK Biobank. EBioMedicine 82: 104161. 10.1016/j.ebiom.2022.104161

[14] Yu X, Rong S (2023) Genome-Wide Associations and Confirmatory Meta-Analyses in Diabetic Retinopathy. Genes (Basel) 14(3). 10.3390/genes14030653

[15] Grassi MA, Tikhomirov A, Ramalingam S, Below JE, Cox NJ, Nicolae DL (2011) Genome-wide meta-analysis for severe diabetic retinopathy. Hum Mol Genet 20(12): 2472-2481. 10.1093/hmg/ddr121

[16] Meng W, Shah KP, Pollack S, et al. (2018) A genome-wide association study suggests new evidence for an association of the NADPH Oxidase 4 (NOX4) gene with severe diabetic retinopathy in type 2 diabetes. Acta Ophthalmol 96(7): e811-e819. 10.1111/aos.13769

[17] Imamura M, Takahashi A, Matsunami M, et al. (2021) Genome-Wide Association Studies Identify Two Novel Loci Conferring Susceptibility to Diabetic Retinopathy in Japanese Patients with Type 2 Diabetes. Hum Mol Genet. 10.1093/hmg/ddab044

[18] Fu YP, Hallman DM, Gonzalez VH, et al. (2010) Identification of Diabetic Retinopathy Genes through a Genome-Wide Association Study among Mexican-Americans from Starr County, Texas. J Ophthalmol 2010. 10.1155/2010/861291

[19] Kim J, Jensen A, Ko S, et al. (2022) Systematic Heritability and Heritability Enrichment Analysis for Diabetes Complications in UK Biobank and ACCORD Studies. Diabetes 71(5): 1137-1148. 10.2337/db21-0839

[20] Guo Y, Yonamine S, Jian Ma C, et al. (2023) Developing and Validating Models to Predict Progression to Proliferative Diabetic Retinopathy. Ophthalmol Sci 3(2): 100276. 10.1016/j.xops.2023.100276

[21] Roden DM, Pulley JM, Basford MA, et al. (2008) Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther 84(3): 362-369. 10.1038/clpt.2008.89

[22] Karlson EW, Boutin NT, Hoffnagle AG, Allen NL (2016) Building the Partners HealthCare Biobank at Partners Personalized Medicine: Informed Consent, Return of Research Results, Recruitment Lessons and Operational Considerations. J Pers Med 6(1). 10.3390/jpm6010002

[23] Wei WQ, Bastarache LA, Carroll RJ, et al. (2017) Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. PLoS One 12(7): e0175508. 10.1371/journal.pone.0175508
[24] Denny JC, Ritchie MD, Basford MA, et al. (2010) PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics 26(9): 1205-1210. 10.1093/bioinformatics/btq126
[25] Carroll RJ, Bastarache L, Denny JC (2014) R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. Bioinformatics 30(16): 2375-2376. 10.1093/bioinformatics/btu197
[26] Sobrin L, Susarla G, Stanwyck L, et al. (2021) Gene Set Enrichment Analyses Identify Pathways Involved in Genetic Risk for Diabetic Retinopathy. Am J Ophthalmol. 10.1016/j.ajo.2021.06.014
[27] Jin H, Jiang D, Ding Z, et al. (2021) Association of four gene polymorphisms in Chinese Guangxi population with diabetic retinopathy in type 2 diabetic patients. BMC Ophthalmol 21(1): 383. 10.1186/s12886-021-02146-4

Table 1. Development and Evaluation of the Electronic Health Record-Based Diabetic Retinopathy Algorithm

Development	Diabetic Retinopathy				Proliferative Diabetic Retinopathy		
	VUMC				VUMC		
	Version 0	Version 1	Version 2	PheCode	Version 0	Version 1	Version 2
Manual Review							
Case (n)	309	220	253	253	86	66	71
Control (n)	125	62	72	72	362*	219*	259*
Undetermined (n)	14	3	5	5	0	0	0
Algorithm							
Case (n)	357	226	271	205	101	59	70
Control (n)	91	59	59	125	347*	226*	260*
PPV	0.857	0.969	0.930	0.922	0.802	0.932	0.914
NPV	0.934	0.966	0.966	0.480	0.986	0.951	0.973

Evaluation	Diabetic Retinopathy			Proliferative Diabetic Retinopathy	
	VUMC n = 330	VANEOHS n = 100	MGB n = 100	VUMC n = 330	VANEOHS n = 121
PPV	0.930	0.940	0.840	0.914	0.818
NPV	0.966	0.860	0.760	0.973	0.989

* Includes individuals with less severe forms of diabetic retinopathy and those without diabetic retinopathy

Table 2: Characteristics of Algorithm-Identified Cohorts in VUMC, MVP, and MGB

Characteristic	VUMC			MVP		MGB	
	DR Cases (N=8,860)	Controls (V0) (N=18,654)	Controls (V1) (N=13,805)	DR Cases (N=62,101)	Controls (N=110,688)	DR Cases (N=471)	Controls (N=1,137)
Mean Age ± SD, years	58.9 ± 14.2	61.3 ± 17.1	62.5 ± 15.3	63.1 ± 9.8	68.5 ± 10.4	67.9 ± 12.4	69.1 ± 13.1
Mean BMI ± SD, kg/m^2	32.0 ± 7.7	32.0 ± 7.7	32.9 ± 7.8	32.5 ± 6.2	32.1 ± 6.5	–	–
Mean Diabetes Duration ± SD, years	6.9 ± 5.3	5.3 ± 4.4	7.4 ± 4.8	6.2 ± 5.1	9.3 ± 5.4	–	–
Female, N (%)	4,441 (50%)	9,979 (53%)	7,274 (53%)	2,648 (4.3%)	7,229 (6.5%)	219 (46%)	512 (45%)
Race/ethnicity							
European American, N (%)	5,912 (67%)	13,857 (73%)	10,176 (74%)	38,507 (62%)	72,999 (66%)	268 (57%)	847 (74%)
African American, N (%)	1,933 (22%)	3,417 (18%)	2,749 (20%)	14,549 (23%)	23,113 (21%)	75 (16%	119 (10%)
Hispanic, N (%)	204 (0.3%)	335 (2%)	247 (1.8%)	6,209 (10%)	9,596 (8.7%)	81 (17%)	32 (2.8%)
Other, N (%)	811 (9.2%)	1,380 (7%)	633 (4.6%)	2,836 (4.6%)	4,980 (4.5%)	47 (10%)	139 (12%)
Type 2 DM, N (%) *	7,437 (84%)	17,556 (94%)	13,239 (96%)	61,429 (99%)	109,724 (99%)	–	–
Genotyped, %	25%	27%	30%	100%	100%	100%	100%
OPT exam in last 5 years, N (%)	2,243 (25%)	12,198 (65%)	4,229 (31%)	44,726 (72%)	78,495 (71%)	–	–
Diabetic macular edema, N (%)	2,167 (25%)	48 (0.3%)	36 (0.3%)	14,180 (22%)	63 (0.05%)	–	–
Neovascular glaucoma, N (%)	501 (5.7%)	195 (1%)	120 (0.9%)	1,723 (2.7%)	1,634 (1.5%)	–	–
Hypertension, N (%)	4,927 (56%)	11,636 (62%)	9,314 (68%)	36,342 (58%)	66,114 (60%)	–	–
Diabetic Neuropathy, N (%)	2,793 (32%)	3,469 (19%)	3,251 (24%)	43,111 (69%)	51,679 (47%)	–	–
Diabetic Nephropathy, N (%)	2,157 (24%)	2,236 (12%)	2,130 (15%)	27,783 (44%)	24,348 (22%)	–	–
Inherited retinal disease, N (%)	54 (0.6%)	156 (0.8%)	92 (0.7%)	459 (0.7%)	607 (0.5%)	–	–
Age-related macular degeneration, N (%)	637 (7.3%)	1,697 (9%)	1,154 (8.4%)	7,678 (12%)	9,622 (6.7%)	–	–

*Type II DM was considered based on a 3:1 Type II to Type I ICD code ratio

Table 3. Diabetes PheCodes Associated with the Diabetic Retinopathy Algorithm Designation

Phenotype	PheCode	Multi n_{max} = 187,581 OR (95% CI)	Non-Hispanic White n_{max} = 128,777 OR (95% CI)	Non-Hispanic Black n_{max} = 43,183 OR (95% CI)	Hispanic n_{max} = 15,621 OR (95% CI)
Diabetes Mellitus	250	2.25 (2.16 – 2.35)	2.17 (2.07 – 2.28)	2.39 (2.21 – 2.58)	68 (33 – 139)
Type I DM (T1D)	250.1	3.48 (3.44 – 3,52)	3.43 (3.38 – 3.48)	3.60 (3.51 – 3.68)	3.56 (3.42 – 3.71)
Type II DM (T2D)	250.2	1.75 (1.66 – 1.85)	1.62 (1.52 – 1.72)	2.12 (1.88 – 2.39)	46.4 (25.8 – 83.5)
T2D with Renal Manifestations	250.22	3.25 (3.22 – 3.29)	3.04 (2.99 – 3.08)	3.63 (3.54 – 3.71)	4.24 (4.08 – 4.41)
T2D with Neurological Manifestations	250.24	2.91 (2.88 – 2.94)	2.71 (2.68 – 2.75)	3.23 (3.16 – 3.30)	3.80 (3.67 – 3.94)
Polyneuropathy in DM	250.6	3.05 (3.02 – 3.09)	2.88 (2.84 – 2.93)	3.41 (3.32 – 3.51)	3.96 (3.78 – 4.14)
Diabetic Retinopathy	250.7	419 (409 – 429)	414 (402 – 426)	383 (365 – 403)	635 (580 – 696)

Table 4. Lessons learned and adaptive strategies in developing case-control algorithms for DR in electronic health records (EHR).

Concern	Lessons Learned/Adaptive strategies
Absence of DR code alone does not guarantee lack of DR in controls.	Require diabetics to have eye exam so that they have had the opportunity to be diagnosed with DR
Access to fundus photography is not uniformly available, disallowing manual review of images. Reliance on structured billing codes may not be accurate for validation	-We encourage review of EHR notes written by providers, and the presence of supporting visits such as eye-exams to allow manual validation with chart review. We further recommend performing sensitivity analyses such as stratifying PPV and NPV by whether diagnoses were provided by eye-specialists. If performance of algorithm is highly skewed where diagnoses made by non-specialists are not reliable, consider using only diagnoses made by specialists. Alternatively, if test properties are comparable, it may be best to maximize available data. -If fundus photography is available for select-patients, review these to make independent determinations of DR diagnoses.
EHR charts are not uniform and often do not have sufficient data for validation	The following strategies may aid in filtering individuals with sufficient information for review: -Require diagnostic codes across multiple visits to confirm diagnosis (greater confidence in determination with reduced sample size) -Use supporting procedural codes such as eye-exam near diagnosis date to increase confidence
Each EHR is unique, and algorithms are not seamlessly portable.	-We discourage automated application of these algorithms. Instead of solely relying on PPVs presented in this paper, we encourage chart review-based validation in a random subset to confirm performance. -We encourage use of these algorithms as a starting point and adapt algorithm to further incorporate any unique features of the new EHR to enhance algorithm performance.
Duration of diabetes is difficult to obtain accurately. **-Severe cases transferred to tertiary facilities may have negative duration of diabetes**	This is a problem that is not unique to EHRs, yet require solutions unique to EHR. -Use earliest date of diagnosis, or anti-diabetic medication mention in records. -If natural language processing capabilities are available, note searches may further help enhance accuracy of these variables. - May need to exclude severe cases with seemingly 'negative' durations. -Create a functional definition of 'incident diabetes' using EHR data and only count cases that occur subsequently. (This will provide the cleanest definition of cases and controls, however, may severely reduce N).
EHR-based case-control studies are not a substitute for epidemiologic studies.	EHR-based case-control studies are powerful complementary resources rather than substitutes for classic epidemiologic studies. -Evidence generated from these highly powered studies can and should be validated in classic epidemiologic designs.
How do I know results obtained from my EHR-based case-control study of DR are comparable to studies obtained from classic epidemiologic designs?	-Carefully consider consequences of choices made in selection criteria. For example, indexing age at last eye visit for controls and indexing age at first DR diagnosis will lead to older controls and possibly longer duration of diabetes recorded in controls, by design, which may be viewed as an adaptive design-approach to control for confounding related to duration of diabetes. -Compare whether sentinel observations from literature hold in your case-control study. For example, do we observe higher proportions of other diabetic complications in cases than in controls, as would be expected.