PLOS ONE

RESEARCH ARTICLE

Discovering novel disease comorbidities using electronic medical records

Shikha Chaganti[1]*, Valerie F. Welty[2], Warren Taylor[3], Kimberly Albert[3], Michelle D. Failla[3], Carissa Cascio[3], Seth Smith[4], Louise Mawn[5], Susan M. Resnick[6], Lori L. Beason-Held[6], Francesca Bagnato[7], Thomas Lasko[8], Jeffrey D. Blume[2], Bennett A. Landman[1]

1 Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, Tennessee, United States of America, **2** Department of Biostatistics, Vanderbilt University, Nashville, Tennessee, United States of America, **3** Department of Psychiatry & Behavioral Sciences, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America, **4** Department of Radiology and Radiological Sciences, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America, **5** Department of Ophthalmology and Visual Sciences, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America, **6** Laboratory of Behavioral Neuroscience, National Institute on Aging, Baltimore, Maryland, United States of America, **7** Department of Neurology, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America, **8** Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America

* cshikha@gmail.com

Citation: Chaganti S, Welty VF, Taylor W, Albert K, Failla MD, Cascio C, et al. (2019) Discovering novel disease comorbidities using electronic medical records. PLoS ONE 14(11): e0225495. https://doi.org/10.1371/journal.pone.0225495

Editor: Vincenzo De Luca, University of Toronto, CANADA

Received: June 7, 2019

Accepted: September 22, 2019

Published: November 27, 2019

Data Availability Statement: The data underlying the results are third party data, and were accessed from the Baltimore Longitudinal Aging Study and Vanderbilt University Medical Center's Synthetic Derivative projects. The authors' data use agreements do not permit dissemination of data. The Synthetic Derivative (SD) is the database containing clinical information derived from Vanderbilt's electronic medical record. DNA samples or genotyping data may be requested after a proposal for the study is received, approved by the BioVU Review Committee and a user agreement is signed. BioVU applications,

Abstract

Increasing reliance on electronic medical records at large medical centers provides unique opportunities to perform population level analyses exploring disease progression and etiology. The massive accumulation of diagnostic, procedure, and laboratory codes in one place has enabled the exploration of co-occurring conditions, their risk factors, and potential prognostic factors. While most of the readily identifiable associations in medical records are (now) well known to the scientific community, there is no doubt many more relationships are still to be uncovered in EMR data. In this paper, we introduce a *novel finding index* to help with that task. This new index uses data mined from real-time PubMed abstracts to indicate the extent to which empirically discovered associations are already known (i.e., present in the scientific literature). Our methods leverage *second-generation p*-values, which better identify associations that are truly clinically meaningful. We illustrate our new method with three examples: Autism Spectrum Disorder, Alzheimer's Disease, and Optic Neuritis. Our results demonstrate wide utility for identifying new associations in EMR data that have the highest priority among the complex web of correlations and causalities. Data scientists and clinicians can work together more effectively to discover novel associations that are both empirically reliable and clinically understudied.

Introduction

Electronic medical record (EMR) systems have been increasingly leveraged for clinical and medical research[1–6]. EMR data provides large sample sizes and information on a wide range

amendments and data use agreements for BioVU and the Synthetic Derivative are tracked through REDCap databases. A detailed description and contact information are at: https://www.vumc.org/dbmi/synthetic-derivative. Additionally, The Baltimore Longitudinal Study on Aging is an NIH intra-mural project. Detailed information on how to apply for data access are at: https://www.blsa.nih.gov.

Funding: LAM: Supported in part by an unrestricted grant to the Vanderbilt Eye Institute and Physician Scientist Award from Research to Prevent Blindness, New York, NY. A dataset used for the analyses described were obtained from Vanderbilt University Medical Center's Synthetic Derivative which is supported by institutional funding and by the Vanderbilt CTSA grant ULTR000445 from NCATS/NIH. BAL: This project was supported National Center for Research Resources, Grant UL1 RR024975-01 (now at the National Center for Advancing Translational Sciences, Grant 2 UL1 TR000445-06). This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University, Nashville, TN. This research was supported by NSF CAREER 1452485 and NIH grants 5R21EY024036. This project was supported in part by ViSE/VICTR. BAL, SR, LBH: This research was conducted with the support from Intramural Research Program, National Institute on Aging, NIH WT: National Institute of Health grant K24 MH110598 and Alzheimer's Association award SAGA-18-418231. SC: National Institute of Biomedical Imaging and Bioengineering training grant T32-EB021937 MF: NIMH training grant T32-MH18921.

Competing interests: The authors have declared that no competing interests exist.

of conditions that impact broad populations. EMR systems contain a rich variety of data including lab results, medications, clinical notes, administrative and billing codes, and images. In this work, we introduce a tool, PheDAS (Phenome-Disease Association Study), to perform association studies and identify disease comorbidities across time in EMR data.

Association studies are typically designed to learn the strength of association between one fixed variable and a large set of potentially correlated variables. An early example was the genome-wide association study (GWAS), which identifies genetic variants associated with a specified phenotypic condition[7]. The phenome-wide association study (PheWAS)[8] reverses the direction to compute the association between many phenotypic conditions and a specific genetic variant. Other studies adopt the design to identify associations between non-genetic variables based on information extracted from EMR. For example, comorbidities of Non-Hodgkin's Lymphoma were found by computing its associations with a set of potential comorbidities extracted from a Medicare claims database[9]. A disease-wide comorbidity map similar to the molecular concept map (MCM) was estimated using an association study design with data extracted from an EMR[10]. Holmes et al used a combination of discharge summaries, diagnostic codes, PubMed database, and Wikipedia articles to identify co-morbidities in three rare diseases[11].

Two of the most important challenges in the design and analysis of association studies are identifying which of the statistical significant results have clinical relevance[12]. An obstacle for traditional statistical methods in these large observational studies is the problem of multiple testing. Because a large number of associations are examined at the same time, the probability of making at least one false claim of significance (the "family-wise error rate") grows with the number of examinations. Previous association studies have employed traditional multiple hypothesis corrections to control either the family-wise error rate or the false discovery rate, such as Bonferroni or Benjamini-Hochberg p-value adjustments[13,14]. Despite these efforts, association studies have suffered from the problem of reproducibility[15]. GWAS studies are usually followed by meta analyses and replication studies to identify truly significant results [16].

Establishing the clinical relevance of results is an important step that is too often missed or done inadequately in high throughput contexts. The most widely used statistical methods ignore clinical relevance; traditional statistical procedures routinely flag significant results that have no practical meaning. When performing inference on such a large scale, it is not feasible to manually sift through all the estimated effects to determine which of the significant results are most clinically important. Besides, this determination ought not to be made after looking at the results. The most common practice is to focus all or most of the attention on the subset of significant results with the smallest p-values. However, findings with the smallest p-values have no guarantee of being the most impactful results, and potential important discoveries are often overlooked. What is desired is a principled way to incorporate clinical relevance into the ranking of important findings.

With the aim of directly addressing these challenges, we used second-generation p-values (p_δ) recently defined by Blume et al to identify associations that are both statistically and clinically significant[17]. Under this approach, we specify *a priori* a null interval hypothesis for effect sizes that are scientifically uninteresting, and only consider as positive findings those associations for which the estimated effect size confidence interval lies completely beyond this null region. Using the second-generation p-value as the metric for defining significant results reduces the type I error and false discovery rates as compared to classical point null hypothesis significance tests which use the p-value[17]. Specifically, incorporating the null interval forces the type I error to zero as the sample size approaches infinity, thereby proactively adjusting the

Type I error to control false discoveries. The subsequent finding are more likely to replicate, almost by definition, and are guaranteed to be clinically relevant.

Exploratory association studies provide an additional third challenge because they can find many associations that are well known, in addition to the potentially surprising results. While the well-known associations can be useful in that they validate the results of the current study and of prior studies, the more important/interesting results are those previously unknown associations which could potentially be used to develop new hypotheses. For external validation of results, previous association studies have used meta analyses,[16] and manual review by experts[11]. Currently there are no methods to identify which associations are empirically reliable but clinically unknown or understudied. In addition, there are no quantitative measures to identify the extent of clinical novelty of these associations. Instead they have to be reviewed on a case-by-case basis which would require a large group of clinical specialists across many areas of expertise and may be more subjective. In this work, we introduce the Novelty Finding Index (NFI), which addresses this challenge and allows for the creation of a tool for mining disease comorbidities that are clinically relevant and can be ranked by novelty (i.e., newness).

Methods

Data

Study 1: Autism Spectrum Disorder. The data for this study, including demographic and ICD-9 codes, was collected at the Vanderbilt University Synthetic Derivative under IRB approval, in a de-identified form. The index disease group for this study was defined by patients with a diagnosis of ASD (ICD-9 codes– 299.*). The control group is defined by subjects with typical development.

Study 2: Alzheimer's disease. The data for this study was collected through the Baltimore Longitudinal Study of Aging (BLSA), a study that collects longitudinal data of an aging population in order to examine changes in the brain as a person ages[18]. The data contains self-reported ICD-9 codes, along with medical records and demographic data. In this study, the index disease group is defined by individuals who were diagnosed with Alzheimer's disease through a clinical consensus. The control group for this study is individuals in BLSA who had no cognitive impairment.

Study 3: Optic neuritis. The data for this study was collected from Vanderbilt University's Synthetic Derivative under IRB approval. It contains EMR data including ICD-9 codes and demographic information. The index disease group is defined by patients with codes 377.30–377.39. The control group for this study are subjects with other disorders of the optic nerve or subjects with hearing loss.

Phenome-disease association study

We developed a python tool to perform phenome-disease association studies (PheDAS). PheDAS is used to identify clinical phenotypes that are associated with a given index disease. A clinical phenotype or a phecode is a code based on hierarchical categorization of ICD-9 (International Classification of Disease—9) codes, which describes a diagnostic "phenotype" by grouping a set of related ICD-9 codes. The ICD-9 codes are coded labels used in billing that describe the relevance of a visit to a particular cluster of symptoms. The visits and procedures themselves are often coded through a separate index known as the Current Procedural Terminology (CPT) system. ICD-9 (or updated -10) codes are associated with each visit or patient history and are necessary for billing to ensure that CPT are codes applicable for each patient. The ~15,000 ICD-9 codes are mapped to 1,865 phecodes as described by Denny et al[19]. For example, "depression" phecode 296.2 groups the ICD-9 codes of "major depressive disorder,

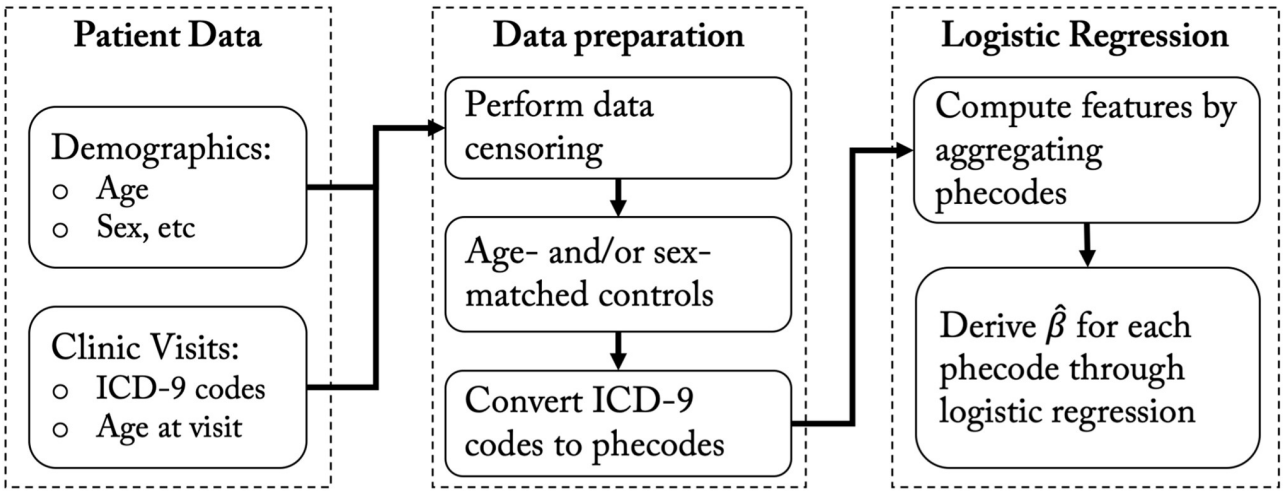

Fig 1. Flow chart for phenome-disease association study. The input patient data required for this analysis is demographic data and clinic visits data. The data is prepared by performing data censoring and control matching based on the experimental design. Next, the ICD-9 codes are converted to phecodes. Finally, logistic regression is performed for each phecode based on aggregate measures and demographic features as described in sections below. To provide a concrete example, consider the Phecode for Alzheimer's Disease (Phecode: 290.11), which references the ICD-9 code 331.0. ICD-9 331.0 maps to CUI C0002395 and then to 41 UMLS strings. Three example strings are "alzheimer's disease", "alzheimers disease", and "senile dementia".

https://doi.org/10.1371/journal.pone.0225495.g001

single episode, mild degree" (ICD-9 = 296.21), "major depressive disorder, recurrent episode, mild degree" (ICD-9 = 292.31), and "depressive disorder NEC" (ICD-9 = 311). For each phecode, a set of exclusion codes are also defined which can be used to select a control cohort.

Given a disease group and a control group, the PheDAS tool performs a set of logistic regressions to identify significant phenotypes associated with the disease. A flowchart of the process is shown in Fig 1. The ICD-9 codes for each clinical visit and other demographic information are extracted from each subject's electronic medical record (EMR). Optionally, the time interval for extraction of ICD-9 codes can also be adjusted according to the study design. This can be done by,

- Censoring by age-interval: Selecting an age range within which to perform the analysis. (Ex. In study 1, we analyze the differences between ASD and control population after age 7); or
- Left-censoring with respect to diagnosis: Selecting a time interval prior to year of diagnosis. (Ex. In study 2, we analyze the differences between Alzheimer's and control population 0–5 years before the diagnosis of the disease); or
- Right-censoring with respect to diagnosis: Selecting a time interval post the year of diagnosis. (Ex. In study 3, we analyze the differences between Optic Neuritis and control population 0–5 years after the diagnosis of the disease).

After defining the interval of the study, ICD-9 codes are extracted for all clinic visits during the period and converted to phecodes using the mapping provided by Denny et al. These codes are denoted by $C = \{c_k | k = 1 \ldots 1{,}865\}$. For each code c_k, an aggregate measure m_k is computed in order to perform the logistic regression. The regression tool can be set to one of the following options:

- Binary measure: aggregate codes to indicate the presence or absence of the phenotype $c_k (m_k = 0 \text{ or } 1)$ in the subject's record in the given time interval,

- Count measure: aggregate codes to indicate the number of times c_k was present in a subject's EMR ($m_k = n$), or
- Duration measure: aggregate codes to indicate the time interval between the first and the last time the phenotype c_k was recorded in a subject's EMR ($m_k = t$).

Additional covariates such as age and sex can also be provided, if available. For each c_k, logistic regression is performed based on the following mean model,

$$\text{logit}(p(class = disease|c_k)) = \beta_0 + \beta_m m_k + \beta_a a_k + \beta_s s_k,$$

where a_k is age and s_k is sex. The coefficient of the aggregate measure β_m is used to determine the significance of the association between the disease and phenotype c_k. In describing the statistical methods, we will denote β_m by θ to allow for some generality, as the methods apply to any parameter of interest in the above regression model. Let the point estimate of $\beta_m = \theta$ be denoted by $\hat{\theta}$.

Second-generation p-value. We used the second-generation p-value (SGPV) measure described by Blume et al[17] to prioritize or rank potential associations. The SGPV framework requires (1) a pre-defined "indifference zone" or null interval hypothesis around the null effect to denote the set of effect magnitudes that would not be clinically meaningful and (2) an uncertainty interval for the observed association, e.g. a confidence interval, likelihood support interval, or credible interval. The SGPV, denoted by p_δ, measures the overlap between the data-supported effect sizes (#2) and the interval null (#1). See Blume (2018) for details[17].

The SGPV equals 0 when #2 and #1 do not overlap. In this case the data only support effect sizes in the alternative hypothesis space. We take all cases where the SGPV is zero, $p_\delta = 0$, to be clinically interesting and statistically 'significant'. In contrast, when $p_\delta = 1$, the data support only effects that are null or nearly null and not of clinical interest. These results would confirm the lack of association. SGPVs between 0 and 1 are treated as inconclusive as the data support both null and alternative hypotheses.

The interpretation of the coefficient θ is different depending on whether m_k is a binary measure, a count measure, or a duration measure. The clinically meaningful effects that make up the null interval will therefore be different in each of these cases. Additionally, the null interval may depend on factors like severity of the outcome. For example, a phecode that increases the odds of having non-specific symptoms such as fever or migraine in Optic Neuritis by a factor of 1.1 may be considered not meaningful, whereas a phecode that increases the odds of having musculoskeletal symptoms in Alzheimer's by a factor of 1.1 could be an important result to consider.

It is of interest to know how reliable SGPV findings are when $p_\delta = 0$ and whether or not the findings are already known in the literature. To address these two important questions, we estimated the positive predictive value (PPV) when the SGPV is zero and developed a "novelty score" by scraping and searching relevant abstracts in PubMed.

Positive predictive value. The positive predictive value (which is the complement of the false discovery rate, FDR) was estimated using an empirical Bayes approach. Define the interval null hypothesis as $H_0 : \theta \in \Theta_0 = [\theta_0^-, \theta_0^+]$ with an alternative hypothesis of $H_1 : \theta \in \Theta_1 = (-\infty, \theta_0^-) \cup (\theta_0^+, \infty)$. Assume that we have a point estimate $\hat{\theta}$ of θ that is asymptotically Normally distributed with variance $\hat{V}_n = V/n$, that is, $\hat{\theta} \sim^A N(\theta, \hat{V}_n)$. The power function for the SGPV is $P(p_\delta = 0|\theta)$ and the form is given in Blume et al.

Under certain assumptions, a reasonable approximation for the PPV is the probability that the null hypothesis is true, given that SGPV equals zero. Applying Bayes' formula, this is $1 - (1 + P(p_\delta = 0|H_1)/P(p_\delta = 0|H_0) \cdot (1 - \pi_0)/\pi_0)^{-1}$ where $\pi_0 = P(H_0)$ is the analysts' a-priori

probability of the null hypothesis before data were collected. We assign probability distributions f_0 and f_1 to the parameter θ under the null and alternative hypotheses, respectively. This allows us to estimate $P(p_\delta = 0|H_1)$ as $1 - \tilde{\beta} = \int_{\Theta_1} P(p_\delta = 0|\theta) f_1(\theta) d\theta$, which is a weighted average of the power function over the alternative space, and estimate $P(p_\delta = 0|H_0)$ as $\tilde{\alpha} = \int_{\Theta_0} P(p_\delta = 0|\theta) f_0(\theta) d\theta$, which is a weighted average of the power function over the null space. Therefore, the PPV is equal to

$$1 - \left[1 + \frac{1 - \tilde{\beta}}{\tilde{\alpha}} \frac{1 - \pi_0}{\pi_0}\right]^{-1},$$

where we set $\pi_0 = 0.5$ which is the default non-informative approach. The probability distribution for the null hypothesis was chosen to be a uniform distribution over the null space, that is, $f_0 \sim Unif[\theta_0^-, \theta_0^+]$. The probability distribution for the alternative hypothesis was chosen to be a uniform distribution over the observed uncertainty interval $(\hat{\theta}_l, \hat{\theta}_u)$, that is, $f_0 \sim Unif[\hat{\theta}_l, \hat{\theta}_u]$. Note that f_1 is a function of the observed data and therefore while the form of f_1 is specified a priori, the actual distribution is not.

Novelty score and novelty finding index. The 'novelty score' is intended to measure the extent to which a finding is well-studied in the literature. We used published abstracts from the PubMed database to construct the 'novelty score' as follows: For each index disease, and for each phecode-disease pairing, we obtained the number of published papers in which these are mentioned in the title, abstract, or keywords section.

In order to search PubMed database, we convert phecodes to search strings using the metathesaurus database provided as a part of the unified medical language system (UMLS) [20], as shown in Fig 2. The UMLS metathesaurus defines unique medical concepts that are unchanged over time, identified by the Concept Unique Identifier (CUI). It links strings with the same meaning from over 200 different source vocabularies to the same CUI. ICD-9 codes are included as a part of source vocabularies provided by UMLS. For each phecode, the ICD-9 codes attached to it are linked to a CUI. Next, all possible strings associated with the CUI are extracted from the metathesaurus to be used as search strings. Henceforth, we will take 'mentioned' to mean the CUI terms linked to a phecode to be mentioned in either the title, abstract, or keywords section.

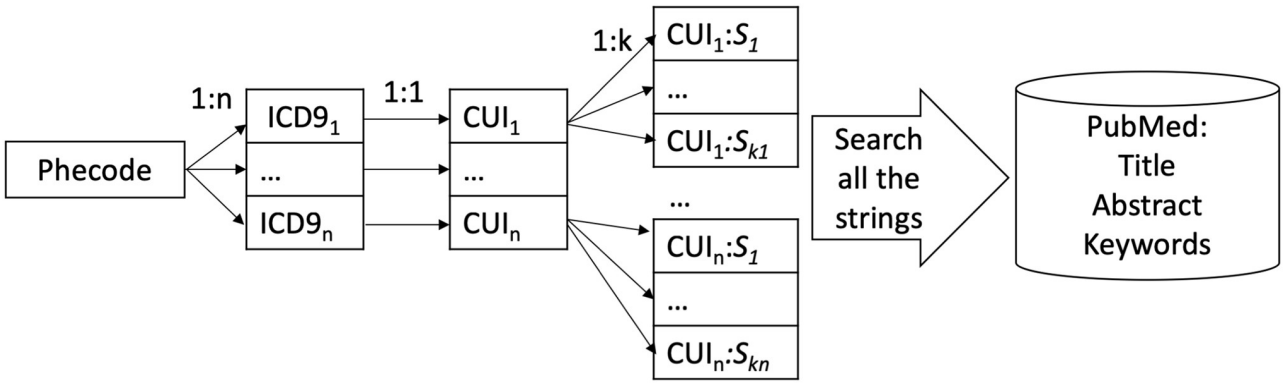

Fig 2. Searching PubMed for associations. For each Phecode, all the ICD-9 codes associated with it are mapped to their CUIs (concept unique identifiers). Next, all the strings associated with the CUIs in UMLS metathesaurus are identified. These strings are used to search all the titles, abstracts and keywords in the PubMed database to identify the counts of academic research papers associated with each phecode.

https://doi.org/10.1371/journal.pone.0225495.g002

We then compute the proportions of published papers that mention the phecode-disease pairing out of all published papers that mention the disease (termed the 'PubMed proportion'). This proportion measures whether the associations between the outcome and the predictor phecodes are well-studied in the literature. Note that well-studied does not necessarily mean well-known to be associated (i.e., the PubMed proportion should not be interpreted as the estimated probability that an association exists). We denoted the novelty score by $N_s = 1 - \hat{F}(x)$, where $\hat{F}(x)$ is the empirical cumulative distribution function estimated with the PubMed proportions under consideration.

Then, to provide a ranking that accounts for both the reliability of the finding (PPV) and its relative novelty (N_s), we define a Novel Finding Index (NFI) as $NFI = (PPV \cdot N_s) \cdot 10$. The purpose of the scale factor of 10 is to move the *NFI* away from the (0, 1) scale, to prevent misinterpretations of the *NFI* as a probability.

Results

PheDAS is an EMR-based open-source association study tool that can be used to evaluate relationships between a fixed condition or disease of interest and other clinical phenotypes. We demonstrate the use of this tool in three studies: 1) Autism Spectrum Disorder (ASD) 2) Alzheimer's Disease, and 3) Optic Neuritis.

The input to the PheDAS tool is in the form of a list of clinical visits with the recorded ICD-9 codes and age at each visit for each subject, along with the status of the condition of interest (0 = absent, 1 = present). For each experiment, a clinically meaningful null-interval is set prior to the analysis. In Figs 3–5, the null interval is indicated by a gray band. The output provides an odds ratio plot for the point and interval estimates, highlighting significant phecodes that are associated with the condition of interest, color-coded by their Novelty Finding Index

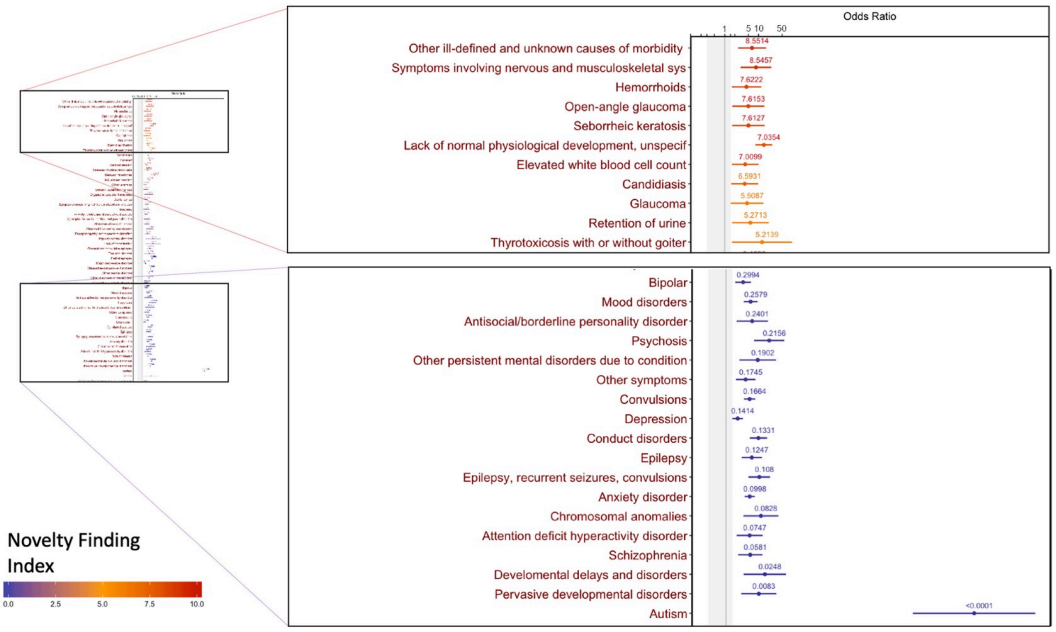

Fig 3. Significant associations for ASD presented as an odds ratio plot. Each dot represents the point estimate of the odds ratio computed from a logistic regression and the line indicates its 95% confidence interval. Each condition is color coded by its novelty finding index, the value of which is displayed above the confidence interval.

https://doi.org/10.1371/journal.pone.0225495.g003

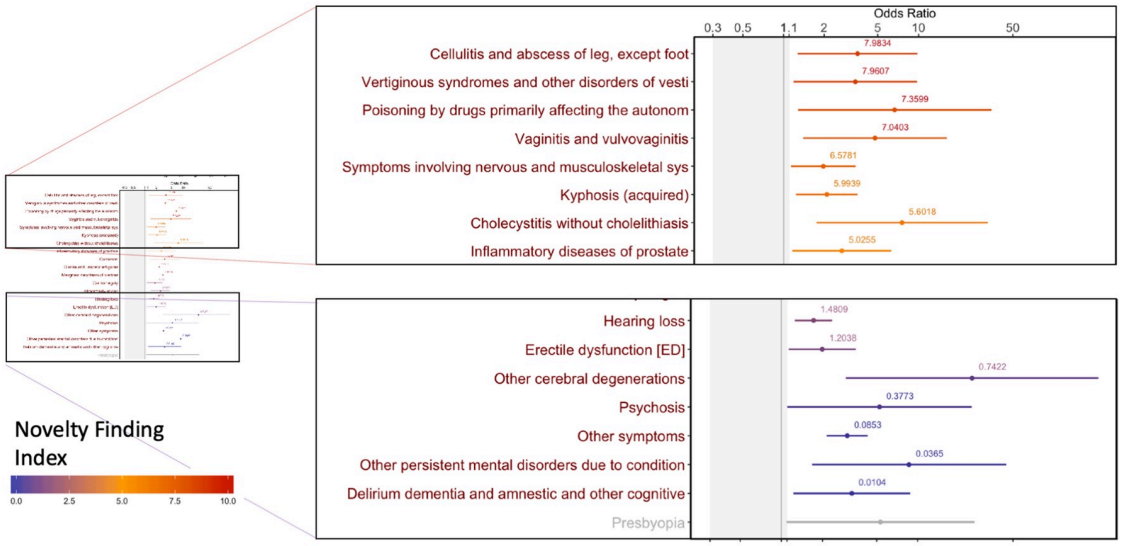

Fig 4. Significant associations 0–5 years before a diagnosis of Alzheimer's disease presented as an odds ratio plot. Each dot represents the point estimate of the odds ratio computed from a logistic regression and the line indicates its 95% confidence interval. Each condition is color coded by its novelty finding index, the value of which is displayed above the confidence interval.

https://doi.org/10.1371/journal.pone.0225495.g004

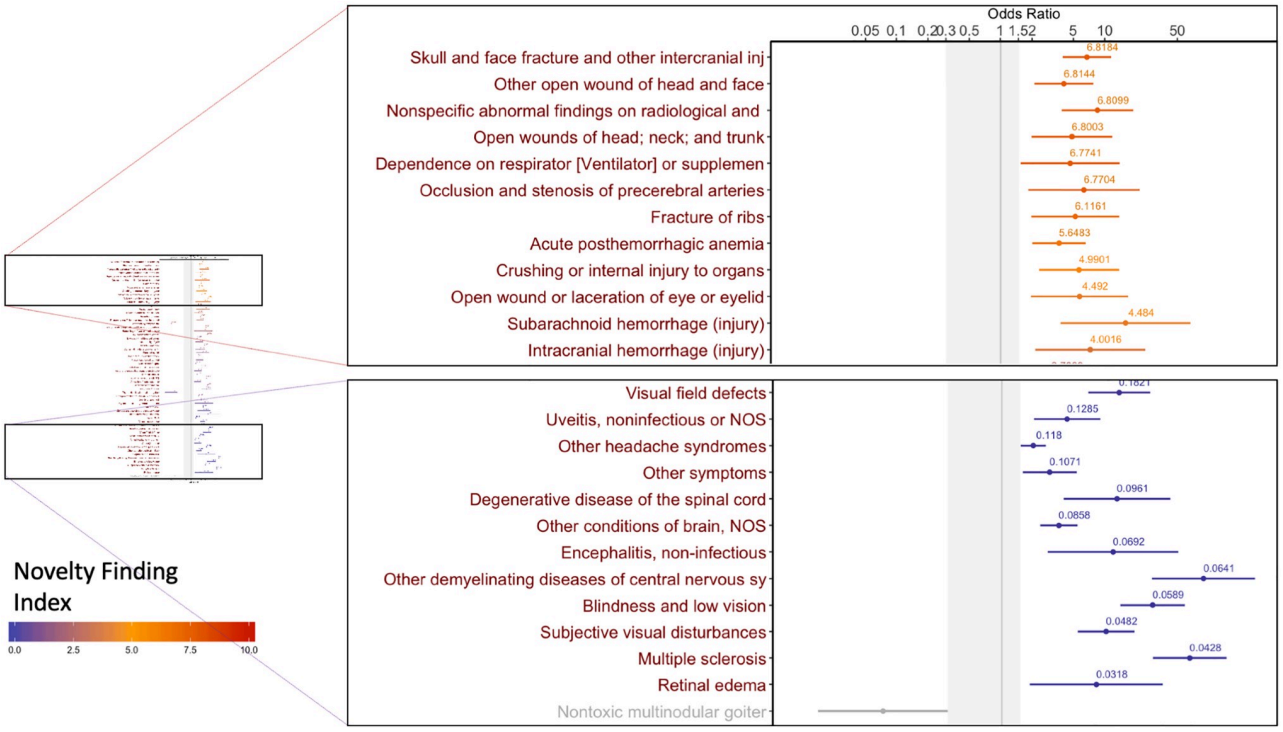

Fig 5. Significant associations 0–5 years after a diagnosis of optic neuritis presented as an odds ratio plot. Each dot represents the point estimate of the odds ratio computed from a logistic regression and the line indicates its 95% confidence interval. Each condition is color coded by its novelty finding index, the value of which is displayed above the confidence interval.

https://doi.org/10.1371/journal.pone.0225495.g005

(NFI). NFI values near 0 suggest that the finding is well known or likely to be a false positive, while values near 10 suggest a novel and reliable finding (Figs 3–5).

Study 1: Autism Spectrum Disorder

We first studied co-morbidities that occur in patients with ASD after early childhood. The dataset included records of 1,234 subjects diagnosed with ASD (926 male and 308 female) and 1,234 age-matched controls without ASD (932 male and 302 female). We included Phecodes for all visits at which the patient was at least seven years old. We chose a null interval of [0.3, 1.5] based on clinician input. In practice, it would be ideal to have the intervals be chosen based on consensus of a team of clinical experts to reduce subjectivity. Note that findings in the "negative" direction (decreased odds of ASD when the phecode is present) were decided to be of less clinical interest than findings in the "positive" direction (increased odds of ASD when the phecode). The same will be true in all 3 examples presented. All data were accessed in de-identified form through Vanderbilt's Synthetic Derivative.

Several strong associations, such as epilepsy, mood disorders, developmental delays, and chromosomal anomalies produced a small NFI, suggesting that these were previously well known (Fig 3).

The novel association of elevated white blood cell count (NFI = 7.01, OR = 4.00, 95% CI 1.62–9.90, SGPV = 0) represents a possible new target to examine a mechanistic theory of ASD. Immune dysregulation has been repeatedly reported in ASD[21] and higher cytokines in response to immune challenge are associated with greater behavioral impairment[22]. Immune dysfunction has clear CNS sequelae, including effects on neurogenesis, synaptic pruning, plasticity, and neuronal function[23]. However, little is known on the mechanistic level about the timing and trajectory of neuro-immune interactions in autism. One theory posits that the role of inflammatory signaling in brain masculinization combined with elevated immune response could explain the high male:female ratio in autism[24]. The potential for addressing these fundamental neurodevelopmental questions using data from electronic health records is stunning, given its longitudinal nature and sample size that allows for the identification of subgroups or the presence of enough females to meaningfully examine sex differences. The novel association detected by PheDAS of elevated leukocytes facilitates more specific hypothesis generation on a topic that has struggled to gain traction in human subjects research. This, for example, could guide future prospective longitudinal studies of neuroimmune interactions in infants at high risk for ASD, moving the field closer to a mechanistic understanding of the impact of immune dysfunction throughout development.

The novel association with glaucoma (NFI = 5.51, OR = 4.54, 95% CI 1.51–13.67, SGPV = 0) may represent another example of how PheDAS may be used. While no current studies have addressed glaucoma in ASD, there is evidence of genetic overlap between glaucoma and ASD[25]. As ASD is a highly heritable disorder, understanding novel genetic associations, especially ones that may be present in unaffected family members can aid early identification and risk assessment for ASD. Additionally, it can highlight the possible common pathways that may represent risk for two different disorders.

Study 2: Alzheimer's disease

We examined predictive factors of Alzheimer's disease by including phecodes of visits between 0 and 5 years prior to the estimated date of diagnosis. The dataset included records of 242 subjects with Alzheimer's Disease (145 male, 97 female) and 789 age- and sex-matched controls with no dementia diagnosis (499 male, 290 female). Matching was performed based on

available data to maximize power. We chose a null interval of [0.3, 1.1] based on clinician input. All data were studied in de-identified form under institutional review board approval.

The well-known associations identified included psychosis, cerebral degenerations, and gait abnormalities (Fig 4). Psychosis and delirium can often be seen in clinical practice and are thought to represent increased risk for neurodegenerative processes including AD.

Novel associations in the five years prior to diagnosis included infections and inflammatory processes across several organ systems. The temporal relationship, wherein the systemic comorbidity precedes clinical diagnosis, supports theories that inflammatory processes and neuroinflammation specifically may contribute to the pathogenesis of AD[26,27]. Peripheral inflammatory markers are elevated early in the AD process [28]and are further associated with cerebrovascular disease[29]. Peripheral elevations in pro-inflammatory cytokines may contribute to neuroinflammation either directly, particularly in situations with compromised blood-brain barrier integrity, or indirectly, through cytokine stimulation of afferent peripheral nerves. Other novel findings include neuromuscular disturbances such as altered vestibular function and increased sensitivity to drugs affecting autonomic function.

Study 3: Optic neuritis

We examined disease progression in optic neuritis by including phecodes of visits between 0 and 5 years after the estimated date of diagnosis, in a population with no previous multiple sclerosis (MS) phecodes. This dataset included 1,085 subjects with optic neuritis (685 male, 405 female) and 1,085 age- and sex-matched controls without a diagnosis of optic neuritis (685 male, 405 female). We chose a null interval of [0.3, 1.5] based on clinician input. All data were studied in de-identified form under institutional review board approval.

The well-known associations identified included visual field defects, subjective visual disturbances, blindness and low vision (Fig 5). These are expected deteriorations of vision from a diseased optic nerve.

A second category of known associations are neurological conditions such as multiple sclerosis, other demyelinating conditions of the brain and degenerative diseases of the spinal cord. These are in line with previous studies that show that there is a significant risk for future MS in patients who have had optic neuritis[30,31].

Novel associations include several conditions related to traumatic injury, such as skull and face fracture (NFI = 6.82, OR = 6.75, 95% CI 3.95–11.55, SGPV = 0), fracture of ribs (NFI = 6.12, OR = 5.22, 95% CI 1.98–13.75, SGPV = 0), crushing or internal injury to organs (NFI = 4.99, OR = 5.67, 95% CI 2.35–13.71, SGPV = 0) and other open wounds of face and neck (NFI = 6.8, OR = 5.75, 95% CI 1.96–16.83, SGPV = 0), suggesting that the partial loss of vision may contribute to such injuries. Acute loss of visual function is well documented in optic neuritis[32]. While visual field defects are recovered within 4 to 7 weeks, there is a delayed mVEP (multifocal visual evoke potential) in regions where there was a visual field loss [33]. An investigation into the relationship between mVEP and increased risk of falls and injuries in optic neuritis population could help explain the novel associations uncovered by PheDAS. Such an investigation can potentially impact therapeutic recommendations for optic neuritis patients to prevent falls and improve visual cognition, such as regular examinations, the use of walk aids, exercise, and video game play[34,35].

An increased risk of fractures has also been noted in patients with MS owing to disability and low bone density[36], so these may also be secondary associations identified by PheDAS. The increased fractures in the immediate aftermath (0–5 yrs) of an optic neuritis diagnosis suggest that some of the symptoms of MS might appear much earlier than previously thought. A longitudinal examination of patients with a diagnosis of optic neuritis and fractures could

reveal new criteria for early prediction of MS. Such a finding can have a substantial impact in management of MS, as early intervention with interferon beta-1b has been shown to delay conversion to clinically definite MS[37,38], reduce the risk for progression of disability[37] and significantly lower lesions on T2-weighted MRI scans[38].

Discussion

In this paper, we describe a new tool for discovering novel disease co-morbidities from routinely-collected EMR data. The co-morbidities may be selectively specified as preceding, co-occurring with, or following the diagnosis of the condition of interest. Our approach can be used for any condition of interest that is captured by the original data collection.

We address the problem of clinical novelty by ranking findings by prior appearance in the scientific literature. We do this by comparing each phecode-disease finding to the number of papers that can be found on PubMed that mention both conditions as a proportion of the number of papers published on the disease of interest. We define a novelty score, which moves the PubMed proportion, which in some sense is on an absolute scale, onto a relative scale. For example, psychosis is scored low on the novelty score (i.e., not considered to be a novel finding) because it is the 4th most frequent predictor phecode that is studied with Alzheimer's, despite the fact that it has only been studied in about 1% of the papers that mention Alzheimer's. A novelty finding index (NFI) is derived from novelty score and the positive predictive value of the association, to indicate the novelty and reliability of the finding.

A limitation of this approach is that we are assuming that if an outcome and predictor phecode are mentioned in the title, abstract, or keywords of the same paper, that an association between them was studied in the paper; this is not necessarily the case. For example, in a paper about Alzheimer's, psychosis may be noted in the 'background' section, but the association between them may not be the topic of the paper. Additionally, papers that study associations among a large number of conditions may not fully list all key terms in the title, abstract, or keywords. These papers would be missed by the proposed approach. However, the fact that the two concepts were discussed in the same paper serves as a reasonable proxy to measure association. As natural language processing (NLP) research advances, it would be interesting to evaluate the of NLP extracts as possible avenues of improving the specificity of patient context[39][40].

NFI has a two-fold benefit. First, the PheDAS methodology can be validated by the novelty finding index by automatically identifying phecode predictors that are well-known by the scientific community. The researcher is assured that the results of the experiment are likely correct, thereby increasing confidence in the analysis. For instance, in our ASD example we see a majority of significant associations that have a low NFI and are well-reported in ASD literature including psychiatric conditions, developmental disorders and seizure disorders. The second advantage, which is the novel aspect of this method is that NFI could be used for hypothesis exploration. It provides a unique and powerful tool to explore empirical relationships in large databases to uncover co-morbidities, risk factors and prognostic factors that were previously under-reported or under-studied. The methodology presented in this paper has the potential to improve understanding of disease etiology and progression and directly impact patient care.

Author Contributions

Conceptualization: Shikha Chaganti, Valerie F. Welty, Warren Taylor, Kimberly Albert, Michelle D. Failla, Carissa Cascio, Seth Smith, Louise Mawn, Susan M. Resnick, Lori L. Beason-Held, Francesca Bagnato, Thomas Lasko, Jeffrey D. Blume, Bennett A. Landman.

Data curation: Shikha Chaganti, Warren Taylor, Susan M. Resnick, Lori L. Beason-Held.

Formal analysis: Shikha Chaganti, Valerie F. Welty, Michelle D. Failla, Jeffrey D. Blume.

Funding acquisition: Louise Mawn, Susan M. Resnick, Lori L. Beason-Held, Bennett A. Landman.

Investigation: Shikha Chaganti, Valerie F. Welty, Warren Taylor, Kimberly Albert, Michelle D. Failla, Carissa Cascio, Seth Smith, Louise Mawn, Susan M. Resnick, Lori L. Beason-Held, Francesca Bagnato, Thomas Lasko, Jeffrey D. Blume, Bennett A. Landman.

Methodology: Shikha Chaganti, Valerie F. Welty, Warren Taylor, Kimberly Albert, Michelle D. Failla, Carissa Cascio, Seth Smith, Louise Mawn, Susan M. Resnick, Lori L. Beason-Held, Francesca Bagnato, Thomas Lasko, Jeffrey D. Blume, Bennett A. Landman.

Project administration: Shikha Chaganti, Louise Mawn, Susan M. Resnick, Lori L. Beason-Held, Thomas Lasko, Bennett A. Landman.

Resources: Michelle D. Failla, Louise Mawn, Susan M. Resnick, Lori L. Beason-Held, Bennett A. Landman.

Software: Shikha Chaganti, Valerie F. Welty, Jeffrey D. Blume.

Supervision: Warren Taylor, Carissa Cascio, Louise Mawn, Susan M. Resnick, Lori L. Beason-Held, Francesca Bagnato, Thomas Lasko, Jeffrey D. Blume, Bennett A. Landman.

Validation: Shikha Chaganti.

Visualization: Valerie F. Welty, Jeffrey D. Blume.

Writing – original draft: Shikha Chaganti, Valerie F. Welty, Warren Taylor, Kimberly Albert, Michelle D. Failla, Carissa Cascio, Thomas Lasko, Jeffrey D. Blume, Bennett A. Landman.

Writing – review & editing: Shikha Chaganti, Valerie F. Welty, Warren Taylor, Kimberly Albert, Michelle D. Failla, Carissa Cascio, Seth Smith, Louise Mawn, Susan M. Resnick, Lori L. Beason-Held, Francesca Bagnato, Thomas Lasko, Jeffrey D. Blume, Bennett A. Landman.

References

1. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Informatics Assoc. 2013; 20: 117–121. https://doi.org/10.1136/amiajnl-2012-001145 PMID: 22955496
2. Coloma PM, Schuemie MJ, Trifirò G, Gini R, Herings R, Hippisley-Cox J, et al. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. Pharmacoepidemiol Drug Saf. 2011; 20: 1–11. https://doi.org/10.1002/pds.2053 PMID: 21182150
3. Richesson RL, Hammond WE, Nahm M, Wixted D, Simon GE, Robinson JG, et al. Electronic health records based phenotyping in next-generation clinical trials: A perspective from the NIH health care systems collaboratory. J Am Med Informatics Assoc. 2013; 20. https://doi.org/10.1136/amiajnl-2013-001926 PMID: 23956018
4. Ahmad NA, Kochman ML, Long WB, Furth EE, Ginsberg GG. Efficacy, safety, and clinical outcomes of endoscopic mucosal resection: a study of 101 cases. Gastrointest Endosc. 2002; 55: 390–396. https://doi.org/10.1067/mge.2002.121881 PMID: 11868015
5. Kellogg TA, Swan T, Leslie DA, Buchwald H, Ikramuddin S. Patterns of readmission and reoperation within 90 days after Roux-en-Y gastric bypass. Surg Obes Relat Dis. 2009; 5: 416–423. https://doi.org/10.1016/j.soard.2009.01.008 PMID: 19540169
6. Turner SD, Berg RL, Linneman JG, Peissig PL, Crawford DC, Denny JC, et al. Knowledge-driven multilocus analysis reveals gene-gene interactions influencing HDL cholesterol level in two independent EMR-linked biobanks. PLoS One. 2011; 6. https://doi.org/10.1371/journal.pone.0019586 PMID: 21589926
7. Manolio TA. Genomewide association studies and assessment of the risk of disease. N Engl J Med. 2010; 363: 166–176. https://doi.org/10.1056/NEJMra0905980 PMID: 20647212

8. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. Bioinformatics. 2010; 26: 1205–1210. https://doi.org/10.1093/bioinformatics/btq126 PMID: 20335276
9. Engels EA, Parsons R, Besson C, Morton LM, Enewold L, Ricker W, et al. Comprehensive evaluation of medical conditions associated with risk of non-Hodgkin lymphoma using Medicare claims ("MedWAS"). Cancer Epidemiol Prev Biomarkers. 2016; cebp-0212.
10. Hanauer DA, Rhodes DR, Chinnaiyan AM. Exploring clinical associations using '-omics' based enrichment analyses. PLoS One. 2009; 4: e5203. https://doi.org/10.1371/journal.pone.0005203 PMID: 19365550
11. Holmes AB, Hawson A, Liu F, Friedman C, Khiabanian H, Rabadan R. Discovering disease associations by integrating electronic clinical data and medical literature. PLoS One. 2011; 6: e21132. https://doi.org/10.1371/journal.pone.0021132 PMID: 21731656
12. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. Am J Hum Genet. 2010; 86: 6–22. https://doi.org/10.1016/j.ajhg.2009.11.017 PMID: 20074509
13. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81: 559–575. https://doi.org/10.1086/519795 PMID: 17701901
14. Browning BL. PRESTO: rapid calculation of order statistic distributions and multiple-testing adjusted P-values via permutation for one and two-stage genetic association studies. BMC Bioinformatics. 2008; 9: 309. https://doi.org/10.1186/1471-2105-9-309 PMID: 18620604
15. Kraft P, Zeggini E, Ioannidis JPA. Replication in genome-wide association studies. Stat Sci A Rev J Inst Math Stat. 2009; 24: 561.
16. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. Am J Hum Genet. 2010; 86: 6–22. https://doi.org/10.1016/j.ajhg.2009.11.017 PMID: 20074509
17. Blume JD, McGowan LD, Dupont WD, Greevy RA Jr. Second-generation p-values: Improved rigor, reproducibility, & transparency in statistical analyses. PLoS One. 2018; 13: e0188299. https://doi.org/10.1371/journal.pone.0188299 PMID: 29565985
18. Resnick SM, Pham DL, Kraut MA, Zonderman AB, Davatzikos C. Longitudinal magnetic resonance imaging studies of older adults: a shrinking brain. J Neurosci. 2003; 23: 3295–3301. https://doi.org/10.1523/JNEUROSCI.23-08-03295.2003 PMID: 12716936
19. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol. 2013; 31: 1102–1111. https://doi.org/10.1038/nbt.2749 PMID: 24270849
20. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings of the AMIA Symposium. American Medical Informatics Association; 2001. p. 17.
21. Mead J, Ashwood P. Evidence supporting an altered immune response in ASD. Immunol Lett. 2015; 163: 49–55. https://doi.org/10.1016/j.imlet.2014.11.006 PMID: 25448709
22. Careaga M, Rogers S, Hansen RL, Amaral DG, Van de Water J, Ashwood P. Immune endophenotypes in children with autism spectrum disorder. Biol Psychiatry. 2017; 81: 434–441. https://doi.org/10.1016/j.biopsych.2015.08.036 PMID: 26493496
23. Meltzer A, Van de Water J. The role of the immune system in autism spectrum disorder. Neuropsychopharmacology. 2017; 42: 284. https://doi.org/10.1038/npp.2016.158 PMID: 27534269
24. McCarthy MM, Wright CL. Convergence of sex differences and the neuroimmune system in autism spectrum disorder. Biol Psychiatry. 2017; 81: 402–410. https://doi.org/10.1016/j.biopsych.2016.10.004 PMID: 27871670
25. Aldinger KA, Lehmann OJ, Hudgins L, Chizhikov VV, Bassuk AG, Ades LC, et al. FOXC1 is required for normal cerebellar development and is a major contributor to chromosome 6p25. 3 Dandy-Walker malformation. Nat Genet. 2009; 41: 1037. https://doi.org/10.1038/ng.422 PMID: 19668217
26. Heppner FL, Ransohoff RM, Becher B. Immune attack: the role of inflammation in Alzheimer disease. Nat Rev Neurosci. 2015; 16: 358. https://doi.org/10.1038/nrn3880 PMID: 25991443
27. Heneka MT, Carson MJ, El Khoury J, Landreth GE, Brosseron F, Feinstein DL, et al. Neuroinflammation in Alzheimer's disease. Lancet Neurol. 2015; 14: 388–405. https://doi.org/10.1016/S1474-4422(15)70016-5 PMID: 25792098
28. King E, O'Brien JT, Donaghy P, Morris C, Barnett N, Olsen K, et al. Peripheral inflammation in prodromal Alzheimer's and Lewy body dementias. J Neurol Neurosurg Psychiatry. 2018; 89: 339–345. https://doi.org/10.1136/jnnp-2017-317134 PMID: 29248892

29. Gu Y, Gutierrez J, Meier IB, Guzman VA, Manly JJ, Schupf N, et al. Circulating inflammatory biomarkers are related to cerebrovascular disease in older adults. Neurol Neuroinflammation. 2019; 6: e521.

30. Group ONS. The 5-year risk of MS after optic neuritis. Experience of the optic neuritis treatment trial. Neurology. 1997; 49: 1404. https://doi.org/10.1212/wnl.49.5.1404 PMID: 9371930

31. Group ONS. Multiple sclerosis risk after optic neuritis: final optic neuritis treatment trial follow-up. Arch Neurol. 2008; 65: 727. https://doi.org/10.1001/archneur.65.6.727 PMID: 18541792

32. Beck RW, Cleary PA, Jye-yu C, Group ONS. The course of visual recovery after optic neuritis: experience of the Optic Neuritis Treatment Trial. Ophthalmology. 1994; 101: 1771–1778. https://doi.org/10.1016/s0161-6420(94)31103-1 PMID: 7800355

33. Hood DC, Odel JG, Zhang X. Tracking the recovery of local optic nerve function after optic neuritis: a multifocal VEP study. Invest Ophthalmol Vis Sci. 2000; 41: 4032–4038. PMID: 11053309

34. Reed-Jones RJ, Solis GR, Lawson KA, Loya AM, Cude-Islas D, Berger CS. Vision and falls: a multidisciplinary review of the contributions of visual impairment to falls among older adults. Maturitas. 2013; 75: 22–28. https://doi.org/10.1016/j.maturitas.2013.01.019 PMID: 23434262

35. Hill K, Schwarz J. Assessment and management of falls in older people. Intern Med J. 2004; 34: 557–564. https://doi.org/10.1111/j.1445-5994.2004.00668.x PMID: 15482269

36. Bazelier MT, van Staa T, Uitdehaag BMJ, Cooper C, Leufkens HGM, Vestergaard P, et al. The risk of fracture in patients with multiple sclerosis: the UK general practice research. J Bone Miner Res. 2011; 26: 2271–2279. https://doi.org/10.1002/jbmr.418 PMID: 21557309

37. Kappos L, Freedman MS, Polman CH, Edan G, Hartung H-P, Miller DH, et al. Effect of early versus delayed interferon beta-1b treatment on disability after a first clinical event suggestive of multiple sclerosis: a 3-year follow-up analysis of the BENEFIT study. Lancet. 2007; 370: 389–397. https://doi.org/10.1016/S0140-6736(07)61194-5 PMID: 17679016

38. Comi G, Filippi M, Barkhof F, Durelli L, Edan G, Fernández O, et al. Effect of early interferon treatment on conversion to definite multiple sclerosis: a randomised study. Lancet. 2001; 357: 1576–1582. https://doi.org/10.1016/s0140-6736(00)04725-5 PMID: 11377645

39. Cheng D, Knox C, Young N, Stothard P, Damaraju S, Wishart DS. PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. Nucleic Acids Res. 2008. https://doi.org/10.1093/nar/gkn296 PMID: 18487273

40. Fiorini N, Canese K, Starchenko G, Kireev E, Kim W, Miller V, et al. Best Match: New relevance search for PubMed. PLoS Biol. 2018. https://doi.org/10.1371/journal.pbio.2005343 PMID: 30153250