Embedding electronic health records onto a knowledge network recognizes prodromal features of multiple sclerosis and predicts diagnosis.

Charlotte A. Nelson[1,4], Riley Bove[2], Atul J. Butte[3,4], Sergio E. Baranzini[1,2,4 *]

[1] Integrated Program in Quantitative Biology. Genentech Hall MC2280. University of California San Francisco. San Francisco, CA. USA

[2] UCSF Weill Institute for Neurosciences. Department of Neurology. University of California San Francisco. 1651 4th St, San Francisco, CA 94158. USA

[3] Department of Pediatrics. University of California San Francisco. 2330 Post St #320, San Francisco, CA 94143. USA

[4] Bakar Computational Health Sciences Institute. University of California San Francisco. 490 Illinois St, Floor 2, Box 2933 San Francisco, CA 94143. USA

* Corresponding author:
Sergio E. Baranzini, PhD
UCSF Weill Institute for Neurosciences. Department of Neurology. University of California San Francisco.
675 Nelson Rising Lane. San Francisco, CA. 94143
Email: Sergio.Baranzini@ucsf.edu
Phone: +1-415-502-6865

Keywords: Knowledge Graph, Electronic Health Records, Multiple Sclerosis, Preventative Medicine

Word Count: 4056

ABSTRACT

Objective

Early identification of chronic diseases is a pillar of precision medicine as it can lead to improved outcomes, reduction of disease burden and lower healthcare costs. Predictions of a patient's health trajectory have been improved through the application of machine learning approaches to electronic health records (EHR). However, these methods have traditionally relied on "black box" algorithms that can process large amounts of data but are unable to incorporate domain knowledge, thus limiting their predictive and explanatory power. Here we present a method for incorporating domain knowledge into clinical classifications by embedding individual patient data into a biomedical knowledge graph.

Materials and Methods

A modified version of the Page rank algorithm was implemented to embed millions of de-identified EHRs into a biomedical knowledge graph (SPOKE). This resulted in high-dimensional, knowledge-guided patient health signatures (i.e. SPOKEsigs) that were subsequently used as features in a random forest environment to classify patients at risk of developing a chronic disease.

Results

Our model predicted disease status of 5,752 subjects three years before being diagnosed with multiple sclerosis (MS) (AUC = 0.83). SPOKEsigs outperformed predictions using EHRs alone, and the biological drivers of the classifiers provided insight into the underpinnings of prodromal MS.

Conclusion

Using data from EHR as input, SPOKEsigs describe patients at both the clinical and biological levels. We provide a clinical use case for detecting MS up to five years prior to their documented diagnosis in the clinic and illustrate the biological features that distinguish the prodromal MS state.

INTRODUCTION

Efforts to move towards precision and preventative medicine have increased in the last decade and are now pervasive in most aspects of biomedicine.[1] As a result, there has been a sharp increase in medical research studies that implement machine learning (ML) approaches using Electronic Health Records (EHRs). [2 3] ML approaches have been moderately successful and have substantially advanced tasks such as disease diagnosis and specimen classification. [4] However, because they identify patterns in data without knowledge of the underlying clinical or biological meaning, their overall performance has been limited and interpretability of the results remains a black box.

Most chronic diseases lack a unique sign or symptom at presentation. On the contrary, patients may consult a specialist following a clinical event, but often acknowledge that symptoms presented months or even years prior. Early identification of individuals at risk for chronic diseases who are still healthy or have subclinical manifestations would be beneficial for both patients (to receive early treatment or close monitoring) and the health system as a whole (to help optimize across multiple visits and expensive testing).

In order to systematically assess the earliest symptoms (i.e. prodromal period) and the biological changes underlying a chronic disease, clinical record standardization is critical in order to overcome the incompleteness in a patient's biomedical history. The Observational Medical Outcomes Partnership (OMOP) format [5] helps bridge the incompatibility of disparate EHR systems and facilitates the unification of patient records and timelines. Additionally, projects that incorporate basic science-level data (genomics, proteomics, etc.) into EHR research, such as Electronic Medical Records and Genomics (eMERGE), have furthered our understanding of disease pathogenesis and offered practical applications [6-8] [9]. A recently recognized need is the consideration of known general biological mechanisms in patient-specific health data analytics. [10] This need can be addressed by knowledge graphs (KG) which naturally bridge the gap between basic science research and medical practice. [11] KGs connect information from multiple classes of biological and medical concepts, thus allowing to constraint the vast solution space faced by traditional ML methods. [12-15] SPOKE is a KG that connects information from over 30 databases and contains more than 3 million nodes of 16 types and more than 16 million edges of 32 types. [16 17] The subset of nodes and edges used here are listed in Tables 1 and 2.

Early detection of chronic diseases such as diabetes or hypertension has enabled their effective management to avoid or delay clinical complications. [18 19] However, despite current efforts in quantifying genetic and environmental risk factors, [20] accurate methods to predict diagnosis of multiple sclerosis (MS) do not yet exist. MS is a chronic, autoimmune disease of the central nervous system (CNS) with severe and life-long consequences. Early symptoms of MS, such as fatigue or depression, are often non-specific, which can make it difficult for the general practitioner to identify and refer the patient to a neurologist. However, previous studies suggest that health care utilization by some patients increases even 10 years prior to their MS diagnosis. [21] Since early treatment of MS is associated with improved long-term neurological outcomes, [22] early recognition of a (sub)clinical presentation and understanding its biological basis could have a major impact on disease trajectories of individual patients. Here we present a computational method to identify patients before they are diagnosed with MS using only the structured portion of their medical records and biological knowledge from a KG. This method for incorporating biological knowledge in health data analysis has broad applicability to other chronic conditions.

MATERIALS AND METHODS

Patient encounter snapshots

The initial cohort consisted of de-identified EHR from 2,180,882 patients who visited UCSF between 2011-2018. Available "snapshots" from the medical history of 5,752 patients with a confirmed diagnosis of MS were taken using only past encounters 1-7 years prior to their first MS diagnosis code (t_0; Figure 1A). These snapshots represent everything a doctor knows about a patient (through their EHRs), up to a given point in time (i.e. snapshot at year -1 contains data up to 1 year before MS diagnosis). These snapshots represent the de-facto prodromal period of MS.

A control group (non-MS, n=2,175,130) was selected among individuals who never received an MS diagnosis during the observational period. For the non-MS group, t_0 was set at 6 months prior to their most recent visit to UCSF. This aligned MS and non-MS snapshots and ensured that the control population had a follow-up period without MS equal to the minimum amount of observation time available for MS patients after diagnosis.

Parallel analyses were conducted to simulate two possible scenarios: patients who visited multiple specialists (*All-Visits*) and patients with only primary or emergency care visits (*PCP-Only*). A patient could potentially be in both simulations if they received both primary and specialist care at UCSF, but only data collected during primary care type visits were used for the *PCP-Only* analysis. Figure 1B depicts the number of MS and Non-MS patients included in the *All-Visits* (left) and *PCP-Only* (right) groups for each snapshot (years -1 to -7).

Embedding EHRs into SPOKE

The EHRs used for this analysis were translated into the OMOP Common Data Model (CDM). We first created Propagated SPOKE Entry Vectors (PSEVs), machine-readable embeddings that quantify the significance of each node in SPOKE for a given cohort of patients. [23] To create PSEVs, SPOKE Entry Points (SEPs) were first identified by finding all concepts that are present in both the EHRs and SPOKE. For this work, we identified 7,535 SEPs, defined as the EHR concepts from the primary tables "condition_occurrence", "drug_exposure", and "measurement" that directly corresponded to nodes in SPOKE. Then, for a given concept (e.g. carbamazepine), a connection was made between a patient SEPs in the EHRs and SPOKE. A modified version of topic-sensitive Page Rank [24] was then used to generate PSEVs for each SEP (Figure 2A-B). Specifically, a random walker was placed onto a node in SPOKE and

allowed it to randomly traverse edges within the network until the walker is forced to restart (p=0.1) at one of the input patients (that was prescribed carbamazepine in this example). This process continues until the amount of time (importance) the walker spends on each node becomes stable. The resulting PSEV holds weights for each node in SPOKE based on how important a node is for the corresponding patient population.

Once population-level embeddings (PSEVs) were created for all matching EHR concepts, they were aggregated to create vectors for the individual patient snapshots. Similar to other machine-learning algorithms, [25 26] we applied vector/matrix arithmetic to produce the Patient Specific SPOKE Profile Vectors (SPOKEsigs, see supplementary Methods). Following this principle, SPOKEsigs were computed for each patient, at each snapshot (Figure 2C). The resulting vectors represent the importance each node in SPOKE for each patient at that time point.

Building a classifier for early detection of MS

Random forest classifiers were used to determine if SPOKEsigs could predict prodromal MS. Random forest was chosen based on its combination of interpretability and performance [27]. To measure the importance of the knowledge network in the prediction, we also created a classifier using only the binary vector corresponding to the patient's SEP. Since SEPs are simply the EHR input variables used to derive the SPOKEsigs, comparing the performance between the two classifiers allowed us to gauge the predictive performance gained by using SPOKE.

In order to build a classifier that could be used to compute risk of MS in the general population, the classifier was tested using the prevalence of MS at UCSF, which approached ~1:1000 for all groups (comparable to the prevalence of MS in the US). [28 29] The classifiers (using either SPOKEsigs or SEPs) were run from snapshots at years -5, -3, and -1 from diagnosis for both the All-Visits and PCP-Only groups.

RESULTS

MS-related nodes increase in significance as time of diagnosis approaches

In order to measure the flow of information from thousands of subjects through the "MS" node in SPOKE before diagnosis, we generated SPOKEsigs without using the PSEV corresponding to the concept *MS* (as *MS* is naturally the top ranked node within the MS PSEV). [23] Of interest, nodes related to the physiopathology of MS were found to be highly ranked likely due to the biologically meaningful connections within SPOKE. To investigate the importance of the *MS* node in our subject population, the rank distribution of *MS* was compared for years -7 to -1 relative to MS diagnosis in the index group. Figure 3a shows that *MS* increases in significance as time to diagnosis approaches for both the *All-Visits* and *PCP-Only* groups (r^2= 0.93; p<0.037 *PCP-Only* and r^2= 0.96; p<0.018 *All-Visits*). Furthermore, when compared to all other diseases in SPOKE, *MS* remains within the top 1% in the *All-Visits* group and (and within 2% for *PCP-Only* visits), during years -7 to -1. Further, the importance of *MS* is statistically significant (T-test) for both groups between years -5 (5.5e-6 *PCP-Only*; 1.6e-26 *All-Visits*) to -1 (6.4e-62 *PCP-Only*; 3.4e-147 *All-Visits*). Note that this cannot be explained by prescriptions of MS-specific disease-modifying medications (DMTs), as these individuals have not been yet diagnosed with MS. There is a noticeable gap between the p-values for the *All-Visits* and the *PCP-Only* groups, suggesting a substantial increase in information related to MS being recorded during specialist visits. Though this increase in significance (overtime as well as the difference between *PCP-Only* and *All-Visits* groups) can partially be attributed to the increased sample size, the average p-value at any time point is not significant. Further, the slope for the MS node compared to the slope of the average p-value over time is 215x and 127x higher (*All-Visits* and *PCP-Only* respectively), suggesting that only a small portion of the increase in significance can be attributed to increased sample size.

To ensure that these results were MS-specific and not simply the outcome of visiting a neurologist (in the *All-Visits* group), a similar analysis was conducted using snapshots from patients diagnosed with Amyotrophic Lateral Sclerosis (ALS). Similarly, *ALS* was the most important disease (p<3.17e-9) in the ALS snapshots at year -1. In contrast, the *MS* node was not differentially ranked compared to the control population (p>0.9). This indicates that although both MS and ALS patients can see neurologists during the prodromal period, each prodromal disease has a distinct signal in SPOKE.

Considering that a first demyelinating event must occur prior to the diagnosis of MS, [30 31] we speculated that SPOKE nodes related to myelin might also increase in significance as time to diagnosis approached. Figure 3b illustrates the increased significance of the concept *Myelin sheath adaxonal region* (GO:0035749). Furthermore, the same trend is observed for any node with "myelin" in its name (Figure 3c). These results suggest that the biological underpinnings of the disease might be detectable during the prodromal period using only information from the EHR.

Predicting Prodromal MS

After confirming that SPOKEsigs contained meaningful information related to MS, a predictive model was built using patient-specific SPOKEsigs as inputs to a random forest classifier. The average AUC for the SPOKEsig All-Visit (AV) classifier was 0.76 at -7 years, and progressively increased to 0.84 for year -1. This same trend was observed for all four classifier types ($AUC^{SPOKE\ AV}$: 0.76-0.84, $AUC^{SPOKE\ PCP}$: 0.6-0.78, $AUC^{SEP\ av}$: 0.7-0.83, $AUC^{SEP\ PCP}$: 0.53-0.75; Figure 4). As expected, the classifier that used all encounters outperformed the classifier that used PCP-Only encounters (Avg. ΔAUC^{SPOKE} Years -1 to -5: 0.11 and Avg. ΔAUC^{SEP} Years -1 to -5: 0.15; Avg. ΔAUC = Avg. AUC All-Visits – Avg. AUC PCP-Only). In all cases of information loss, either from smaller time windows (time from diagnosis) or missing specialist visits (PCP-Only), the enhancement of EHRs with SPOKE drove classifier performance. The greatest improvement was seen at three years prior to diagnosis using PCP-Only encounters ($\Delta AUC^{SPOKE-SEP}$: 0.12). Altogether, these results demonstrate that embeddings of patients' clinical data from the structured portion of the EHR onto a KG contain relevant information about their health status. Furthermore, adding structured knowledge to EHR data through SPOKE can compensate for missing and incomplete EHR data.

More SEPs will likely improve classifier performance

We recognize SEPs themselves are incomplete because they currently do not map every EHR concept to SPOKE (88% of conditions, 79% of medications, and 47% of measurements for *All-Visits* at year -1). To estimate how much SPOKEsigs could improve if each EHR concept was mapped to SPOKE, the same classifiers were run using the full set of EHR concepts. Interestingly, the average difference in AUC between full OMOP and SPOKEsigs was the same as that between SPOKEsigs and SEPs (ΔAUC: 0.053). The majority of OMOP concepts that drove the full OMOP classifiers were measurements that were not mapped to SPOKE

(Supplementary Tables 1 and 2). These results suggest that if more EHR concepts were mapped to SPOKE, a significant improvement in the classifier could be achieved.

Biological drivers of the classifier

Our previous results suggest that the improved performance of classifiers using SPOKEsigs over those using only SEPs (i.e. straight from the EHR) is due to biologically relevant information from SPOKE being utilized in the computation (i.e. because the network connects these variables). To understand how the incorporation of biological knowledge increased the AUC, we extracted the scores of each biological node using the average feature value across all years for both the *All-Visits* and *PCP-Only* groups. Next the top 20 nodes from each biological node type (*Gene, Protein, Biological Process, Molecular Function, Cellular Component*, and *Pathway*) were selected and split into MS or Non-MS significant groups according to the sign of the t-statistic (Figure 5 a and b respectively). To further interpret how each group of top nodes were connected to one another, additional SPOKE nodes were added if they had direct edges to at least two top biological nodes (Figure 5a-b). Remarkably, the highest ranked nodes in the MS groups corresponded to myelin biology (*myelin sheath adaxonal region*, *MAG, glial cell differentiation* etc.), neurophysiological functions (*axonogenesis, ceramide binding*, etc.) and adaptive immunity (*CD4+T cells and B cell-specific pathways, CCR5*, etc.) (Figure 5a, Supplementary Table 3). Also significant were nodes related to the CNS, muscle behavior, the extracellular matrix (e.g. *matrix metalloproteinases, collagen, NCAM, Basigin interactions*, etc.), and genes associated with other neurological diseases such as spastic paraplegia (*MPV17L2*), ataxia (*RNF170*) Alzheimer's disease (*APBA3*), and lysosomal storage disease (*NAGLU*). Together, these nodes illustrate how the classifier detected the importance of neurological and immunological processes in MS patients several years before their diagnoses. In contrast, the highest ranked nodes within the Non-MS group were related to Th2 cell differentiation (*eosinophil migration, prostaglandins, CCR3 chemokine receptor binding*, etc.), an immune subset associated with protection against inflammatory diseases like MS (Figure 5b). [32-35]

Medications and common laboratory tests drive information flow to neurological nodes

The difficulty in identifying MS at an early stage is due to the combination of the EHRs being sparse and MS symptoms being vague and common in the general population. Often this results in OMOP codes only being associated with one or a small number of MS patients (Supplementary Figure 2) which does not contribute to the classifier. However, after mapping an OMOP concept to a SEP it is transformed into a multidimensional SPOKEsig that represents the

importance of each node in SPOKE for that OMOP concept/SEP. Therefore, two distinct OMOP concepts could “push” information to the same downstream nodes.

To identify which OMOP concepts were responsible for “pushing” information downstream to each of the MS-significant biological nodes, network paths were traced back to the originating SEPs (see methods). For most of the top MS nodes, the SEPs that were essential for the high rank of the MS-significant nodes were mapped from medication orders and common laboratory tests (note that MS DMTs are not SEPs, as none of these individuals had been diagnosed with MS at the time of analysis). Though these SEPs may not have been significant in the MS population as a whole, their propagation through SPOKE led to increased information flow to the MS top nodes. For example, while Carbamazepine and Lithium are not significant as distinct SEPs, they both direct information flow to the GO concept “Myelin sheath adaxonal region” (GO: 0035749, a highly ranked MS-relevant node) in a representative patient shown in Figure 5c. For this patient, information flows from Carbamazepine to a set of Disease nodes (either through “treated by” or “contraindicated for” edges) and then (either directly or through an additional *Disease* or *Gene* node) to the genes CNP, MAG, or PTEN which are all components of “Myelin sheath adaxonal region”. Interestingly, Carbamazepine or Lithium can be used to treat symptoms and comorbidities of MS such as trigeminal and glossopharyngeal neuralgia or depression, respectively, which are common symptoms experienced by MS patients. This further demonstrates that distinct clinical presentations can lead to similar SPOKE representations of MS patients.
Similarly, the paths between the laboratory test for Aspartate aminotransferase travel through aspartic acid (Compound) and then traverse one to two edge(s) before reaching MAG and PTEN (Genes) (Supplementary Figure 3). Despite the different paths of entry into SPOKE, data are repetitively sent through nodes such as MAG and PTEN, which then converge at the “Myelin sheath adaxonal region” node. Similar patterns were observed for multiple other neurological nodes.

Th2-mediated diseases drive information to Non-MS biological nodes

The same method for abstracting the pertinent OMOP concepts information flow was then applied to the top Non-MS biological nodes. After retracing several paths, we found that the OMOP concepts that facilitated the flow of information to nodes related to eosinophils, eicosanoids, and T-cells were driven by Th2-mediated diseases such as asthma and allergies

which are more prevalent in the Non-MS population (–log2 odds ratio of -2.46 and -1.97 accordingly). Figure 5d provides an example of how these diseases transfer information to the (non-MS significant) biological node *Eicosanoid ligand-binding receptors*. In this representative patient, data start at the node for *asthma* and then either directly connect to or are one neighbor apart from genes that participate in *Eicosanoid ligand-binding receptor* (*Pathway*). In the latter case, the information first flows through diseases similar to *asthma* or its associated genes. These straightforward routes from Th2-mediated diseases to their associated genes are what power the Th2 signal in the Non-MS significant biological nodes.

Taken together, our results show that SPOKE nodes useful for the classifier include nodes with both strongly positive (highly ranked in MS) and negative (highly ranked in controls) associations with MS. In both cases, the biological interpretation of those nodes is consistent with the known pathogenesis of MS.

DISCUSSION

The purported prodromal period of MS is often described in terms of health care utilization. [36 37] MS patients in the prodromal stage are, by definition, months or even years away from a recorded diagnosis code for MS. During this period, however, they are not just standing idly - in fact, their healthcare use both within and beyond the primary care setting, steadily increases until time to diagnosis. [36] Previous research revealed that MS patients have more encounters with psychiatrists and urologists, as well as higher proportions of musculoskeletal, genito-urinary or hormonal-related prescriptions. [38]These findings hint that underlying biological signals must be present months or even years before diagnosis and the information from these specialist visits could be pivotal in uncovering those differences.

While patients often pay multiple visits to a specialist before receiving an MS diagnosis, the process of obtaining an appointment with a specialist can itself be prolonged, usually requiring a referral and insurance coverage. As a result, a patient's initial interface with a health system is often through primary or emergency care. Appreciating the different roles primary care and specialist clinicians play in the diagnosis process, we ran two analyses in parallel using data from either primary care providers only (*PCP-Only*) or all visit types (*All-Visits*). Though it is possible for symptoms to be recorded in the structured portion of EHRs, this typically only occurs if it is necessary for billing. Additional patient data can be extracted from the patient notes using natural language processing (NLP). However, NLP methods to date generate rather sparse data, and need further validation in healthcare settings; thus their incorporation is out of scope for this work.

The generation of PSEVs is comparable to word2vec, another machine-learning vector embedding method. [25 26] Similar to how word2vec learns the embedding of a word by using the words around it as context, PSEVs utilize patient cohorts to give context to the nodes in SPOKE. PSEVs are then added together to produce the Patient Specific SPOKE Profile Vectors (SPOKEsigs) that describe a patient in terms of node weights in SPOKE. The main difference between these two embedding techniques is that PSEVs (and therefore SPOKEsigs) are based on a "clear box" algorithm that constructs machine-readable vectors while maintaining human interpretability. This means each element in the vector corresponds to a node in SPOKE and it is possible to trace back how information travels from sparse EHRs to downstream nodes. The diffusion of EHRs through SPOKE enabled the prioritization of the *MS Disease* node in the

SPOKEsigs of MS patients compared to controls. Additionally, the significance of this differential prioritization increases as the time to diagnosis decreases. Further, we have shown that the known biological underpinnings of MS could be abstracted using these sparse clinical data. This is evident by the prioritization of myelin related nodes within the SPOKEsigs of MS patients – whose disease is characterized by demyelination in the CNS - compared to controls up to seven years prior to MS diagnosis.

We hypothesized that SPOKEsigs contained deeper information about a patient than the equivalent EHR vectors (SEPs). Remarkably, SPOKEsigs outperformed SEPs (i.e. EHR-only information) at all time points for both the *All-Visits* and *PCP-Only* analyses. The *All-Visit* AUCs were always higher than the *PCP-Only* AUCs due to the greater power of the *All-Visit* group in both number of patients and encounters. This difference was minimized by the addition of SPOKE, which enabled the use of *PCP-Only* data to achieve results closer to using *All-Visit* data using the SEPs alone. This enhancement of EHRs using SPOKE was particularly striking for the *PCP-Only* analysis performed 3 years before diagnosis, which showed a 12% improvement in AUC (over SEPs alone). These results hint at a future where, after adequate validation including consideration of possible biases, SPOKE could be used at the point of care to support or target supplementary evaluation for primary care providers.

The top biological drivers of the classifier were split into two groups (MS significant or Non-MS significant) according to whether they were ranked higher in the MS or Control SPOKEsigs. Notably, neurophysiological functions, CNS, and muscle behavior nodes were among the top MS-significant nodes. In contrast, there were many Th2-related nodes (indicating immunoregulatory activity) dominating the Non-MS significant nodes. Interestingly, *phospholipase C activity*, which was high in the MS group, is known to play a role or interact with in both the MS and Non-MS top immune features. Moreover, phospholipase C [39] was recently implicated in female-specific neuropathic pain induce a myelin basic protein peptide (MBP_{84-104}) in mice. This study showed that after MBP exposure, T-cells attack the DRG and spinal cord in females but remain localized in males. [40] Notably multiple top nodes from both the MS and Non-MS groups participate together in this pathway in a way that is consistent with both this observed sexual dimorphism as well as the increased prevalence of MS among women. This connection between top immune nodes within MS and Non-MS groups further supports the hypothesis that MS (and others like RA) results from an imbalance between proinflammatory (Th1 or Th17) and immunoregulatory Th2 responses. [41] In contrast, asthma

and allergies are mediated by Th2 responses, which presumably protect against Th1/Th17-driven diseases. [42 43]

PSEVs represent a new class of clear (as opposed to a black) box algorithms. This property allowed us to trace back how key biological nodes became significant. The propagation of information to nodes that were ranked higher in non-MS patients mostly originated from Th2 mediated diseases such as allergies and asthma, which were more prevalent in the non-MS population. In contrast, a heterogeneous set EHRs mainly from commonly ordered laboratory tests or treatments for comorbidities facilitated information to move to the MS significant nodes. These results demonstrate that clinical presentation and biological changes are inherently linked and the intersection can be uncovered using EHRs during the MS prodromal period.

To move towards the delivery of precision medicine, disease biology and clinical manifestations must be investigated side by side. Increasing amounts of data are being obtained for individual patients, and knowledge networks will play a key role in bridging the gap between biological knowledge derived from basic science research, and medical knowledge. As more measurements (genomics, proteomics, microbiome) become available, we hypothesize the SPOKEsigs will become even more informative. Further, the transition from curative to preventative medicine can only be possible through a better understanding of the prodromal biology of a disease. It is our hope that such methods will be used for a variety of diseases to advance both precision and preventative medicine.

CONCLUSIONS:

This work presents a strategy to embed EHR data onto a knowledge graph (SPOKE) to obtain high-dimensional health status profiles (SPOKEsigs). SPOKEsigs were computed for hundreds of thousands of individuals and a random-forest classifier was trained to identify individuals at risk of MS. This approach was able to detect MS up to five years prior to their documented diagnosis in the clinic. SPOKEsigs represent a new kind of “clear box” explainable predictable models with broad applicability to other chronic medical conditions where early diagnosis can benefit patients.

Data Sharing Statement

Due to the sensitive nature of EHR, we are not able to share patient data, even in de-identified form. To facilitate the reproducibility and advancement of this research, we have created an API for generating SPOKEsigs alongside a jupyter notebook with instructions on how to use it, which can be accessed at https://github.com/BaranziniLab/SPOKEsigs. Anyone with access to EHRs can now create SPOKEsigs for their own patient populations and test the concepts presented in this work. SPOKE can be accessed at https://spoke.rbvi.ucsf.edu/neighborhood.html.

FIGURE LEGENDS

Figure 1. Patient timeline aligning and filtering. (a) Timepoint 0 (t_0) is the point of alignment for the MS and Non-MS timelines. For MS patients, t_0 was the first visit in which a patient received a diagnosis code for MS. The duration of time a patient has been diagnosed with MS is represented by a red line between the first and last visits with a MS diagnosis code. For Non-MS patients t_0 was set to 6 months (purple line) prior to their most recent visit (hexagon). Left of t_0 are the patient snapshots that encompass all of the information (EHR data) a doctor has on a patient up to a given point of time. The snapshot at year -1 (blue line) contains all data between the first visit (triangle) and -1 year from t_0. The remaining snapshots (years -3, -5, and -7) become smaller as their endpoints move farther from t_0. **(b)** Two patient encounter groups were followed throughout the workflow: *All-Visit* (left) and Primary Care Physician (*PCP-Only*) (right). The *All-Visit* analysis uses all possible encounters at UCSF, while the *PCP-Only* analysis only includes patient encounters at primary (or emergency) care visits. The number of MS or Non-MS patients at each year go from t_0 (top) to -7 years (bottom) is shown.

Figure 2. Embedding individual patients in SPOKE. (a) Example embedding the EHR concept for the drug carbamazepine into SPOKE. First, SPOKE Entry Points (SEPs) are created by finding all concepts that are present in both the EHRs and SPOKE. Then each patient that was prescribed carbamazepine is connected to SPOKE through the SEPs in their EHRs. A random walker is then placed onto a node in SPOKE and randomly traverses edges within the network until the walker is restarted at one of the patients that was prescribed carbamazepine (probability of restart = 0.1). **(b)** This process continues until the amount of time the walker spends on each node becomes stable. The nodes are then ranked such that the most important nodes are given the highest rank (dark teal) and the least important nodes are given the lowest rank (white). Here the medically or biologically important nodes for carbamazepine are darker teal. Meanwhile, *heartburn*, which is not related to carbamazepine, is white. **(c)** A SPOKEsig is produced for a patient at a given snapshot by summing the PSEVs associated with the SEPs in their EHRs during that time period. During this example snapshot, Patient X had three SEPs: *carbamazepine, epilepsy*, and *constipation*. Therefore, the PSEVs for *carbamazepine, epilepsy*, and *constipation* are summed to create this snapshot for Patient X.

Just as the elements in the PSEVs, each element in the SPOKEsig corresponds to a single node in SPOKE.

Figure 3. MS biology nodes become more significant with time to diagnosis. (a-c) For each node, the t-test was used to compare the distribution of ranks between the MS and Non-MS patients. Here the -log10 P-value from the t-test is plotted against time to MS diagnosis (or lack thereof) for the **(a)** *MS* node (DOID:2377), **(b)** *Myelin sheath adaxonal region* (GO:0035749), and **(c)** group of nodes with "myelin" in the name.

Figure 4. Integrating SPOKE enhances classifier AUC. ROC curves for predicting MS diagnosis at year(s) -1, -3, and -5 (**a-c** accordingly) with a random forest classifier. The classifiers that used encounters from All-Visits are in blue (SPOKEsig input vector) and green (SEP input vector). The classifiers that only used encounters from PCP visits are shown in orange (SPOKEsig input vector) and red (SEP input vector). In all instances the SPOKEsig input vectors out preformed the corresponding SEP input vector. The largest gain in AUC was for the PCP encounter classifier 3 years prior to diagnosis.

Figure 5. Th1/Th2 balance and neurological nodes drive biological increase in AUC. **a-b** Networks of significant biological nodes for random forest classifier. Red nodes were higher ranked in the MS population **(a)**, while blue nodes were higher ranked higher in the non-MS population **(b)** (color gradient based on t-statistic). The shape (diamond or oval) of the node denotes whether or not the node is in the top 20 of a given node type. If it is an oval, it must connect to >= 2 nodes in the top 20. Highlighted in the network are some of the nodes that correspond to th1/th2 balance or neurology. **(c)** Illustration of how a prescription for carbamazepine can send information to the *Myelin sheath adaxonal region node* (GO:0007404, GO:0043360; replaced by: GO:0010001). **(d)** Depiction of how *asthma* (a th2 mediated disease that is more prevalent in the non-MS UCSF population) pushes information downstream to the *Eicosanoid ligand-binding receptors* node (Reactome R-HSA-391903).

TABLES

Table 1. SPOKE node statistics	
Node Type	Count
Compound	286790
Protein	33857
Gene	19567
Anatomy	13257
BiologicalProcess	13156
Disease	9128
SideEffect	3865
MolecularFunction	3407
Pathway	2428
PharmacologicClass	1748
CellularComponent	1725
Symptom	369

Table 2. SPOKE edge statistics			
Node Type 1	Edge	Node Type 2	Count
Disease	ASSOCIATES_DaG	Gene	1998072
Gene	PARTICIPATES_GpBP	BiologicalProcess	1480742
Protein	INTERACTS_PiP	Protein	1238535
Compound	BINDS_CbP	Protein	1098776
Anatomy	EXPRESSES_AeG	Gene	1052814
Gene	REGULATES_GrG	Gene	531344
Gene	INTERACTS_GiG	Gene	294328
Gene	PARTICIPATES_GpMF	MolecularFunction	260152
Gene	PARTICIPATES_GpCC	CellularComponent	226582
Gene	PARTICIPATES_GpPW	Pathway	221080
Anatomy	DOWNREGULATES_AdG	Gene	204480
Anatomy	UPREGULATES_AuG	Gene	195696
Disease	RESEMBLES_DrD	Disease	128000
Gene	COVARIES_GcG	Gene	123380
Compound	CAUSES_CcSE	SideEffect	86400
Disease	LOCALIZES_DlA	Anatomy	79010
Protein	TRANSLATEDFROM_PtG	Gene	67332
Compound	TREATS_CtD	Disease	64872
PharmacologicClass	INCLUDES_PCiC	Compound	62952
Disease	PRESENTS_DpS	Symptom	47606
Compound	DOWNREGULATES_CdG	Gene	42204
Compound	CONTRAINDICATES_CcD	Disease	41302
Compound	UPREGULATES_CuG	Gene	37512
Anatomy	ISA_AiA	Anatomy	37304
Anatomy	CONTAINS_AcA	Anatomy	37304
Disease	ISA_DiD	Disease	22952
Disease	CONTAINS_DcD	Disease	22952
Anatomy	PARTOF_ApA	Anatomy	19502
Disease	UPREGULATES_DuG	Gene	15462
Disease	DOWNREGULATES_DdG	Gene	15246
Compound	RESEMBLES_CrC	Compound	12972
Compound	INTERACTS_CiP	Protein	6390
Compound	PALLIATES_CpD	Disease	780
Compound	AFFECTS_CamG	Gene	718

Disclosure Statements:

Funding Statement: This work was supported in part by a grant from the US National Science Foundation (Convergence Accelerator NSF_1937160). Partial support also comes from the Bakar Family Foundation and the Bakar Computational Health Sciences Institute. SEB is holds the Heidrich Family and Friends endowed chair in Neurology at UCSF. SEB holds the Professorship in Neurology I at UCSF.

Competing Interests Statement:

RB is funded by the National Multiple Sclerosis Society Harry Weaver Award, the National Science Foundation and the National Institutes of Health National Library Medicine.
She has received research grant funding from Biogen, Novartis and Roche Genentech
She has received consulting honoraria from Alexion, Biogen, EMD Serono, Genzyme Sanofi, Novartis, and Roche Genentech. S.E.B. is co-founder and holds shares in MATE Bioservices, a company that commercializes uses of SPOKE knowledge graph. C.A.N holds shares of MATE Bioservices. A.B is a co-founder and consultant to Personalis and NuMedii; consultant to Samsung, Mango Tree Corporation, and in the recent past, 10x Genomics, Helix, Pathway Genomics, and Verinata (Illumina); has served on paid advisory panels or boards for Geisinger Health, Regenstrief Institute, Gerson Lehman Group, AlphaSights, Covance, Novartis, Genentech, Merck, and Roche; is a shareholder in Personalis and NuMedii; is a minor shareholder in Apple, Facebook, Alphabet (Google), Microsoft, Amazon, Snap, Snowflake, 10x Genomics, Illumina, Nuna Health, Assay Depot (Scientist.com), Vet24seven, Regeneron, Sanofi, Royalty Pharma, Pfizer, BioNTech, AstraZeneca, Moderna, Biogen, Twist Bioscience, Pacific Biosciences, Editas Medicine, Invitae, Doximity, and Sutro, and several other non-health related companies and mutual funds; and has received honoraria and travel reimbursement for invited talks from Johnson and Johnson, Roche, Genentech, Pfizer, Merck, Lilly, Takeda, Varian, Mars, Siemens, Optum, Abbott, Celgene, AstraZeneca, AbbVie, Westat, several investment and venture capital firms, and many academic institutions, medical or disease specific foundations and associations, and health systems. A.B receives royalty payments through Stanford University, for several patents and other disclosures licensed to NuMedii and Personalis. A.B's research has been funded by NIH, Northrup Grumman (as the prime on an NIH contract), Genentech, Johnson and Johnson, FDA, Robert Wood Johnson Foundation, Leon Lowenstein Foundation, Intervalien Foundation, Priscilla Chan and Mark Zuckerberg, the Barbara and Gerson Bakar Foundation, and in the recent past, the March of Dimes, Juvenile

Diabetes Research Foundation, California Governor's Office of Planning and Research, California Institute for Regenerative Medicine, L'Oreal, and Progenity.

Contributorship Statement:

CAN: Developed methods for analysis, gathered data, and performed analysis. Created Figures, and drafted manuscript,

RB: Analyzed clinical data, ensured medical accuracy, edited manuscript.

AJB: Contributed to the study design, methods development and manuscript revision and editing.

SEB: Study conception, design and supervision. Data analysis and Figure design. Drafted and edited manuscript. Accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

REFERENCES

1. Auffray C, Charron D, Hood L. Predictive, preventive, personalized and participatory medicine: back to the future. Genome medicine 2010;**2**(8):57 doi: 10.1186/gm178[published Online First: Epub Date]|.
2. Tate AR, Beloff N, Al-Radwan B, et al. Exploiting the potential of large databases of electronic health records for research using rapid search algorithms and an intuitive query interface. J Am Med Inform Assoc 2014;**21**(2):292-8 doi: 10.1136/amiajnl-2013-001847[published Online First: Epub Date]|.
3. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012;**13**(6):395-405 doi: 10.1038/nrg3208[published Online First: Epub Date]|.
4. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE J Biomed Health Inform 2018;**22**(5):1589-604 doi: 10.1109/JBHI.2017.2767063[published Online First: Epub Date]|.
5. Stang PE, Ryan PB, Racoosin JA, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Ann Intern Med 2010;**153**(9):600-6 doi: 10.7326/0003-4819-153-9-201011020-00010[published Online First: Epub Date]|.
6. Denny JC, Ritchie MD, Basford MA, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics 2010;**26**(9):1205-10 doi: 10.1093/bioinformatics/btq126[published Online First: Epub Date]|.
7. Ritchie MD, Denny JC, Crawford DC, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. Am J Hum Genet 2010;**86**(4):560-72 doi: 10.1016/j.ajhg.2010.03.003[published Online First: Epub Date]|.
8. Kho AN, Pacheco JA, Peissig PL, et al. Electronic medical records for genetic research: results of the eMERGE consortium. Sci Transl Med 2011;**3**(79):79re1 doi: 10.1126/scitranslmed.3001807[published Online First: Epub Date]|.
9. Consortium e. Harmonizing Clinical Sequencing and Interpretation for the eMERGE III Network. Am J Hum Genet 2019;**105**(3):588-605 doi: 10.1016/j.ajhg.2019.07.018[published Online First: Epub Date]|.
10. Gustafsson M, Nestor CE, Zhang H, et al. Modules, networks and systems medicine for understanding disease and aiding diagnosis. Genome Medicine 2014;**6** doi: ARTN 82 10.1186/s13073-014-0082-6[published Online First: Epub Date]|.

11. Nicholson DN, Greene CS. Constructing knowledge graphs and their biomedical applications. Computational and Structural Biotechnology Journal 2020;**18**:1414–28 doi: 10.1016/j.csbj.2020.05.017[published Online First: Epub Date]|.
12. Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network–based approach to human disease. Nature reviews. Genetics 2011;**12**(1):56–68 doi: 10.1038/nrg2918[published Online First: Epub Date]|.
13. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. The human disease network. Proc Natl Acad Sci U S A 2007;**104**(21):8685–90 doi: 0701361104 [pii]
10.1073/pnas.0701361104[published Online First: Epub Date]|.
14. Wang L, Himmelstein DS, Santaniello A, Parvin M, Baranzini SE. iCTNet2: integrating heterogeneous biological interactions to understand complex traits. F1000Research 2015 doi: 10.12688/f1000research.6836.1[published Online First: Epub Date]|.
15. Zhou X, Menche J, Barabasi AL, Sharma A. Human symptoms–disease network. Nat Commun 2014;**5**:4212 doi: 10.1038/ncomms5212[published Online First: Epub Date]|.
16. Himmelstein DS, Baranzini SE. Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease–Associated Genes. PLoS Comput Biol 2015;**11**(7):e1004259 doi: 10.1371/journal.pcbi.1004259[published Online First: Epub Date]|.
17. Woolhandler S, Himmelstein DU. Single–Payer Reform. Ann Intern Med 2017;**167**(7):527 doi: 10.7326/L17–0326[published Online First: Epub Date]|.
18. OConnor PJ, Sperl–Hillen JM, Rush WA, et al. Impact of electronic health record clinical decision support on diabetes care: a randomized trial. Ann Fam Med 2011;**9**(1):12–21 doi: 10.1370/afm.1196[published Online First: Epub Date]|.
19. Ye C, Fu T, Hao S, et al. Prediction of Incident Hypertension Within the Next Year: Prospective Study Using Statewide Electronic Health Records and Machine Learning. Journal of Medical Internet Research 2018;**20**(1) doi: 10.2196/jmir.9268[published Online First: Epub Date]|.
20. Xia Z, Steele SU, Bakshi A, et al. Assessment of Early Evidence of Multiple Sclerosis in a Prospective Study of Asymptomatic High–Risk Family Members. JAMA Neurol 2017;**74**(3):293–300 doi: 10.1001/jamaneurol.2016.5056[published Online First: Epub Date]|.
21. Disanto G, Zecca C, MacLachlan S, et al. Prodromal symptoms of multiple sclerosis in primary care. Ann Neurol 2018;**83**(6):1162–73 doi: 10.1002/ana.25247[published Online First: Epub Date]|.

22. Kappos L, Edan G, Freedman MS, et al. The 11-year long-term follow-up study from the randomized BENEFIT CIS trial. Neurology 2016;**87**(10):978-87 doi: 10.1212/WNL.0000000000003078[published Online First: Epub Date]|.
23. Nelson CA, Butte AJ, Baranzini SE. Integrating biomedical research and electronic health records to create knowledge-based biologically meaningful machine-readable embeddings. Nat Commun 2019;**10**(1):3045 doi: 10.1038/s41467-019-11069-0[published Online First: Epub Date]|.
24. Topic-sensitive pagerank. Proceedings of the 11th international conference on World Wide Web; 2002. ACM.
25. Mikolov T, Chen K, Corrado G, J. D. Efficient estimation of word representations in vector space. 2013;(arXiv:1301.3781v3).
26. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. Lake Tahoe, Nevada: Curran Associates Inc., 2013:3111-19.
27. Couronne R, Probst P, Boulesteix AL. Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics 2018;**19**(1):270 doi: 10.1186/s12859-018-2264-5[published Online First: Epub Date]|.
28. Hirtz D, Thurman DJ, Gwinn-Hardy K, Mohamed M, Chaudhuri AR, Zalutsky R. How common are the "common" neurologic disorders? Neurology 2007;**68**(5):326-37 doi: 10.1212/01.wnl.0000252807.38124.a3[published Online First: Epub Date]|.
29. Wallin MT, Culpepper WJ, Campbell JD, et al. The prevalence of MS in the United States. Neurology 2019;**92**(10):e1029-e40 doi: 10.1212/wnl.0000000000007035[published Online First: Epub Date]|.
30. Okuda DT, Mowry EM, Beheshtian A, et al. Incidental MRI anomalies suggestive of multiple sclerosis The radiologically isolated syndrome. Neurology 2009;**72**(9):800-05 doi: 10.1212/01.wnl.0000335764.14513.1a[published Online First: Epub Date]|.
31. Okuda DT, Siva A, Kantarci O, et al. Radiologically isolated syndrome: 5-year risk for an initial clinical event. PLoS One 2014;**9**(3):e90509 doi: 10.1371/journal.pone.0090509[published Online First: Epub Date]|.
32. Raphael I, Nalawade S, Eagar TN, Forsthuber TG. T cell subsets and their signature cytokines in autoimmune and inflammatory diseases. Cytokine 2015;**74**(1):5-17 doi: 10.1016/j.cyto.2014.09.011[published Online First: Epub Date]|.

33. Hirahara K, Nakayama T. CD4+T-cell subsets in inflammatory diseases: beyond the Th1/Th2 paradigm. International Immunology 2016;**28**(4):163-71 doi: 10.1093/intimm/dxw006[published Online First: Epub Date]|.
34. Bomprezzi R. Dimethyl fumarate in the treatment of relapsing-remitting multiple sclerosis: an overview. Ther Adv Neurol Disord 2015;**8**(1):20-30 doi: 10.1177/1756285614564152[published Online First: Epub Date]|.
35. Crane IJ, Forrester JV. Th1 and Th2 Lymphocytes in Autoimmune Disease. Crit Rev Immunol 2005;**25**(2):75-102 doi: 10.1615/CritRevImmunol.v25.i2.10[published Online First: Epub Date]|.
36. Wijnands JM, Kingwell E, Zhu F, et al. Infection-related health care utilization among people with and without multiple sclerosis. Mult Scler 2016:1352458516681198 doi: 10.1177/1352458516681198[published Online First: Epub Date]|.
37. Wijnands JMA, Kingwell E, Zhu F, et al. Health-care use before a first demyelinating event suggestive of a multiple sclerosis prodrome: a matched cohort study. The Lancet Neurology 2017;**16**(6):445-51 doi: 10.1016/s1474-4422(17)30076-5[published Online First: Epub Date]|.
38. Wijnands JMA, Zhu F, Kingwell E, et al. Five years before multiple sclerosis onset: Phenotyping the prodrome. Multiple Sclerosis Journal 2018;**25**(8):1092-101 doi: 10.1177/1352458518783662[published Online First: Epub Date]|.
39. Bush WS, McCauley JL, DeJager PL, et al. A knowledge-driven interaction analysis reveals potential neurodegenerative mechanism of multiple sclerosis susceptibility. Genes & Immunity 2011;**12**(5):335-40 doi: 10.1038/gene.2011.3[published Online First: Epub Date]|.
40. Chernov AV, Hullugundi SK, Eddinger KA, et al. A myelin basic protein fragment induces sexually dimorphic transcriptome signatures of neuropathic pain in mice. Journal of Biological Chemistry 2020;**295**(31):10807-21 doi: 10.1074/jbc.RA120.013696[published Online First: Epub Date]|.
41. Bar-Or A. The Immunology of Multiple Sclerosis. Seminars in Neurology 2008;**28**(1):029-45 doi: 10.1055/s-2007-1019124[published Online First: Epub Date]|.
42. Tremlett HL. Asthma and multiple sclerosis: an inverse association in a case-control general practice population. Qjm 2002;**95**(11):753-56 doi: 10.1093/qjmed/95.11.753[published Online First: Epub Date]|.
43. Eagar TN, Miller SD. Helper T-Cell Subsets and Control of the Inflammatory Response. Clinical Immunology, 2019:235-45.e1.

44. Spackman KA, Campbell KE, Cote RA. SNOMED RT: a reference terminology for health care. Proc AMIA Annu Fall Symp 1997:640–4
45. Schriml LM, Arze C, Nadendla S, et al. Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res 2012;**40**(Database issue):D940–6 doi: 10.1093/nar/gkr972[published Online First: Epub Date]|.
46. Kibbe WA, Arze C, Felix V, et al. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic acids research 2014;**43**(D1):D1071–D78
47. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;**32**(Database issue):D267–70 doi: 10.1093/nar/gkh061[published Online First: Epub Date]|.
48. Wishart DS, Knox C, Guo AC, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res 2006;**34**(Database issue):D668–72 doi: 10.1093/nar/gkj067[published Online First: Epub Date]|.
49. Law V, Knox C, Djoumbou Y, et al. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res 2014;**42**(Database issue):D1091–7 doi: 10.1093/nar/gkt1068[published Online First: Epub Date]|.
50. Gaulton A, Hersey A, Nowotka M, et al. The ChEMBL database in 2017. Nucleic Acids Res 2017;**45**(D1):D945–D54 doi: 10.1093/nar/gkw1074[published Online First: Epub Date]|.
51. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. Journal of the American Medical Informatics Association 2011;**18**(4):441–48 doi: 10.1136/amiajnl–2011–000116[published Online First: Epub Date]|.
52. Degtyarenko K, de Matos P, Ennis M, et al. ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Research 2007;**36**(Database):D344–D50 doi: 10.1093/nar/gkm791[published Online First: Epub Date]|.
53. McDonald CJ, Huff SM, Suico JG, et al. LOINC, a universal standard for identifying laboratory observations: a 5–year update. Clin Chem 2003;**49**(4):624–33
54. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene–centered information at NCBI. Nucleic Acids Res 2011;**39**(Database issue):D52–7 doi: 10.1093/nar/gkq1237[published Online First: Epub Date]|.
55. Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. Uberon, an integrative multi–species anatomy ontology. Genome Biology 2012;**13**(1) doi: 10.1186/gb–2012–13–1–r5[published Online First: Epub Date]|.

56. Ursu O, Holmes J, Knockel J, et al. DrugCentral: online drug compendium. Nucleic Acids Res 2017;**45**(D1):D932–D39 doi: 10.1093/nar/gkw993[published Online First: Epub Date]|.
57. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Research 2016;**44**(D1):D1075–D79 doi: 10.1093/nar/gkv1075[published Online First: Epub Date]|.
58. Gene Ontology C. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research 2004;**32**(90001):258D–61 doi: 10.1093/nar/gkh036[published Online First: Epub Date]|.
59. Xu Q–S, Liang Y–Z. Monte Carlo cross validation. Chemometrics and Intelligent Laboratory Systems 2001;**56**(1):1–11 doi: 10.1016/s0169–7439(00)00122–2[published Online First: Epub Date]|.
60. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and regression trees*: Routledge, 2017.
61. Strobl C, Boulesteix A–L, Zeileis A, Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 2007;**8**(1) doi: 10.1186/1471–2105–8–25[published Online First: Epub Date]|.

SUPPLEMENTARY

SPOKE

SPOKE is a continuously updated biomedical knowledge network including over 5 decades of research and the version frozen for this analysis consists of 389,297 nodes (including genes, diseases, symptoms, compounds, etc.) and 9,774,753 edges, derived from high throughput research databases. Tables 1 and 2 detail the node and edge types in the version of SPOKE used for this analysis. By navigating the network (either manually or computationally) a user can identify connecting points between any two concepts, and even more powerful, merge with additional information to derive new knowledge. SPOKE can be accessed through the Neighborhood Explorer tool https://spoke.rbvi.ucsf.edu/neighborhood.html.

Aligning patient timelines and filtering patients

We first identified patients who received at least one diagnosis code for MS (MS patient group) and those who never received a diagnosis code for MS (Non-MS patient group). The MS patient population was subsequently filtered to only keep those with an MS diagnosis for >= 6 months (between the first MS diagnosis code (t_0) and the last diagnosis code) and >=5 MS-related encounters (unique dates that a patient visited UCSF and a MS diagnosis code was documented in their record). Additionally, MS patients who were prescribed an MS disease-modifying therapy prior to an MS diagnosis code were removed from the population because it was not possible to obtain an accurate diagnosis date for those patients.

In order to align the Non-MS patient timelines with the MS patient timelines, the Non-MS patients were required to have a matched Non-MS diagnosis observation period of 6 months. As a result, t_0 for Non-MS patients is set at 6 months prior to their most recent encounter.

Once t_0 was established for both the MS and Non-MS groups, we created snapshots of patients up to 7 years prior to t_0. These snapshots aim to represent all that a clinician has learned about a given patient from the first time the patient visited UCSF until their visit at year -1, -3, -5, or -7 years from diagnosis. Seven years prior to diagnosis is the farthest we can go back because the current UCSF EHR system started in 2011.

A final filter was placed to remove patients with too little information during the usable encounter period. For the statistical analysis, a light filter was applied that required patients have at least three OMOP concepts and three SEPs. This filter was made more stringent for the

classification, requiring at least five OMOP and SEPs. For the classifier, MS patients also had to visit UCSF and receive an MS diagnosis code at least 5 times. These stringent filters resulted in a reduced number of patients (Supplementary Table 4) and a prevalence of on average ~1:1000 (0.15% and 0.11% for *All-Visit*s and PCP-Only respectfully).

Dividing encounter types

Since most patients interact with non-specialists (primary care, emergency, family medicine, etc.) more frequently than specialists, our analysis was carried out in parallel using two encounter groups. The first encounter group (*All-Visits*) used EHR data from any encounter without regard for a clinician's specialty. The second group (Primary Care Provider Only or *PCP-Only*) only used encounters from Internal Medicine, Hospital Medicine, Family Medicine, Emergency Medicine, Urgent Care, or General Practice.

It should be noted that, though in practice it is true that patients would interact with the *PCP-Only* group more than the *All-Visits* group, UCSF is a specialty-focused institution with a comparatively limited primary care division. Therefore, the majority of patients within the UCSF EHR system only see specialists at UCSF. This is apparent by the size of the *PCP-Only* patient population size (Figure 1b), which is approximately one fifth the size of the *All Visit* patient population.

Translating OMOP concepts to SEPs

UCSF EHR data up to October 2018 were transformed to use the OMOP CDM. The tables used were condition_occurrence, drug_exposure, visit_occurrence, provider, and measurement. The visit_occurrence and provider tables facilitated the categorization of encounters into *All-Visits* or *PCP-Only*. The remaining tables were then mapped to nodes in SPOKE to create SPOKE Entry Points (SEPs). Since OMOP utilizes standard terminologies, mapping between OMOP and SPOKE was greatly accelerated compared to previous efforts [23]. The UCSF condition_occurrence concepts used the vocabulary SNOMED [44] that were mapped to *Disease* (DiseaseOntology ID [45 46]), *Symptom* (MeSH ID), or *SideEffect* (Unified Medical Language System [47] UMLS CUI) SPOKE nodes using relationships in DiseaseOntology and UMLS. The concepts in the drug_exposure table (RxNorm vocabulary) were mapped to *Compound* (DrugBank [48 49] and/or ChEMBL IDs [50]) nodes. These mappings were accomplished using tables from RxNorm [51] and Chemical Entities of Biological Interest [52](ChEBI). Finally, concepts from the measurement table (LOINC ID [53]) were mapped to a variety of nodes (*Compounds, Genes [54], Anatomies [55], Diseases, PharmacologicalClasses*

[56], SideEffects [57], and *GeneOntologies* [58]) through UMLS relationships. In order to translate LOINC to SPOKE, the UMLS CUI could be translated a maximum of two times before mapping to a SPOKE node. Additionally, the translations were filtered by relationship (RELA) to avoid one-to-many mappings. The translations to *Compound* nodes utilized the concept_class_id in OMOP to distinguish the drug classes (e.g. ingredient vs brand name) for more specific translations.

Creation of SPOKEsigs

PSEVs were created for each SEP as previously described (Nelson et al, 2019; Figure 2a). Next, Patient Specific SPOKE Profile Vectors (SPOKEsigs) were generated for each patient at each defined snapshot. This was achieved by summing the PSEVs that were associated with the SEPs within an individual patient's snapshot (usable encounters). All of the nodes were then ranked (from 1 to the number of nodes in SPOKE) where the most important node was equal to the number of nodes in SPOKE ($SPOKEsig^{inital\ rank}$). To highlight the nodes that were the most important for each patient the matrix of SPOKEsigs was z-score normalized ($SPOKEsig^{z\text{-}score}$). Finally, to enhance the biological heterogeneous nature of the SPOKEsigs, the nodes were ranked by node type ($SPOKEsig^{rank\ by\ type}$) for each patient. Again, the most important node was given the largest value (i.e. the number of nodes of a given type).

Odd and p-value calculations

To access differences in the EHR records between the MS and Non-MS populations at different snapshots, a confusion matrix was produced for each OMOP concept or SEP. For laboratory tests, patients had to have an abnormal result for the measurement concept to be counted. The confusion matrix for a given concept or SEP was then used as input for the Fisher's exact test (python package: scipy.stats.fisher_exact) to generate the odds ratio and p-value.

T-stats for significant nodes

PSEVs can be generated for any EHR concept. Nodes that are biologically or medically important for a given EHR concept will be prioritized within the PSEV [23]. To see if this held true after aggregation for individual patients ($SPOKEsig^{initial\ rank}$), we used the t-test to derive p-values and t-statistics to compare the distribution of ranks of the MS *Disease* node in the MS and Non-MS populations at years -1, -3, -5, and -7 years from diagnosis. This was repeated for other nodes known to be integral to the biology of MS such as myelin related nodes.

Random Forest Classifier

The RandomForestClassifier from the sklearn python package was used for the OMOP, SEP, and SPOKEsig random forests. Due to the size of the SPOKEsigs, dimensionality was reduced to increase efficiency and respect memory restrictions. The number of nodes used as input was reduced to only include the most variant (>= 50th overall percentile) and top ranked (<=20,000 by type) nodes on average for the entire population. For *All-Visits* (or *PCP-Only*), this resulted in 60,150 (40,700), 52,963 (40,570), and 53,171 (45,198) nodes for year -1, -3, and -5 respectively. The number of nodes seen in all classifiers was 43,677 for *All-Visits* and 36,956 for *PCP-Only*. Year -7 was dropped in this part of the analysis due to a low number of MS patients. Additionally, bootstrapping was used to limit the number of patients (n=10,000) used to train each base estimator. For each run, 20% of the patients were held out from the training group to be used for testing. To approach exhaustive cross-validation, we used Monte Carlo cross validation for 10,000 different random splits of patients.[59] To ensure that the results were comparable, same training and testing populations were used for the SPOKEsigs, SEPs and all OMOP classifiers.

Identifying top biological driver nodes for classifier

Within the RandomForestClassifier function is the property feature_importances, which holds the Gini importance [60 61] of each element for the input vector. The Gini importance was averaged across the 10,000 rounds and then each node was ranked by type where the top ranked node is now 1. The rank of each node was then averaged across the years. This process yielded the score of each node for both the *All-Visits* and *PCP-Only* groups. The final score of each node was calculated by summing the score for the *All-Visits* and *PCP-Only* groups. To focus on the most important biological nodes, the top 20 nodes were selected from: *BiologicalProcess, CellularComponent, Pathway, MolecularFunction, Gene,* or *Protein* node types.

Interconnected top biological network

Since most of the top biological nodes (*BiologicalProcess, CellularComponent, Pathway,* and *MolecularFunction*) are from biological systems, the natural way to see if they are related is through the genes they share. Likewise, the remanding nodes (*Genes* and *Proteins*) can be related through shared biological systems, interactions, and co-expression. To illustrate these relationships, the top nodes were connected together in a network with additional biological nodes that had edges to at least two of the top biological nodes. The network was then

separated into networks based on t-statistics where the MS significant network had nodes with a positive t-statistic and Non-MS significant network had nodes with a negative t-statistic. To facilitate visualization, the additional SPOKE biological nodes were filtered in the "Non-MS" and "Overall" networks. These nodes were filtered based on the number of edges they had to top nodes compared to the number of edges they had to any biological node in SPOKE:

$$\frac{Edges_{node\ 1\ in\ top\ biological\ network}}{Edges_{node\ 1\ in\ all\ biological\ network}}$$

Retracing paths from SEP to SPOKE

A score was calculated to determine which SEPs were the most responsible for initiating information flow to the top biological nodes. The score for a top node (TN) took two metrics into consideration: the value of the TN in each PSEV and the odds ratio of a SEP in the population. First, the value of the TN within each PSEV was converted into the percentile, where the PSEV with the highest TN values would be equal to 1 and the lowest equal to 1/number of nodes. Next, the percentile of the TN was multiplied by the –log2 fold change of the associated SEP. Finally, if the average t-statistic of the TN within the SPOKEsigs was positive, (top nodes that are higher in MS) then the highest scored SEP(s) were selected. In contrast, if the average t-statistic of the TN within the SPOKEsigs was negative, (top nodes that are higher in Non-MS), then the lowest scored SEP(s) were selected. Once a top SEP was established, all possible paths of length less that 3 were found between the SEP and TN. The nodes within the paths between SEP and TN were then filtered according to their value in the PSEV of the corresponding SEP.

Investigating sparse EHRs and common MS symptoms

It is known that early MS often presents with vague symptoms that are common in the general population. To investigate this and our most significant results (a difference of 0.12 AUC between SPOKEsigs and SEPs at year -3 using PCP-only) we examined the input OMOP concepts. First, the OMOP data for patients at the year -3 snapshot PCP-only were filtered by removing those that could not be mapped to a SEP. Next, the data was split into MS and Non-MS cohorts. The OMOP concepts were then groups according to whether they were recorded for 1, 2, or >=3 patients. The pie charts in Supplementary Figure 2 (left) show the number of OMOP concepts in each of these groups by OMOP domain: Condition (a), Drug (b), and Measurement (c). These results show that the early EHR codes for MS patients are very sparse.

Next, the log2 fold change (MS compared to Non-MS) was calculated for each OMOP concept and the distribution was displayed for each group using a violin plot (Supplementary Figure 2 right). Here, it is clear that OMOP codes with the highest fold change are associated with a low number of MS patients. This exemplifies the diversity in early MS presentation. Further, the most frequently recorded OMOP concepts in the MS population tend to have a negative log2 fold change. This demonstrates how vague MS symptoms can be and that the common OMOP concepts for MS patients are also common for the control population.
Interestingly, the only domain with a higher proportion of common OMOP concepts was Measurement. This is consistent with our earlier conclusion that improved (more complete and precise) mapping of measurements to SPOKE could improve the performance of the classifier.

Quantifying differences between models

The output to each model included the ROC plot, mean AUC, and standard deviation AUC (mean and standard deviation from the cross-validations). After the models were finished, we evaluated whether the AUCs for the SPOKEsig models were significantly different than those in the SEP models. Given that the AUCs follow a normal distribution, the mean and standard AUC from a model was used to generate a sample AUC distribution. For each year and cohort (All Visit or PCP-Only) the sample SPOKEsig and SEP distributions were compared using the Mann-Whitney U test. Sample distribution generation and p-value calculations repeated 1000 to calculate the average p-value. As expected, the comparison between SPOKEsigs and SEPs at year -3 using PCP-Only has the most significant p-value. It should be restated that these are only approximate p-values based on the models' mean and standard deviations.

Supplementary Figures and Tables

Supplementary Figure 1. Measurements increase accuracy of disease predictions. Not all data can be mapped to SPOKE. Most unmapped EHR concepts are measurements (lab tests), which are the most biologically substantive parts of the EHRs. To evaluate how much SPOKEsigs could be improved once these concepts are mapped the same random forest models were run using all data. **(a-b)** show ROCs for random forest classifier using all OMOP

data for **(a)** *All-Visit*s or **(b)** PCP-Only groups for snapshots at years -1 (red), -3 (orange), and -5 (yellow).

Supplementary Figure 2. EHR sparsity and vague common MS symptoms. (a-c) Pie charts and violin plots describing OMOP code frequency in the PCP-Only year -3 MS population per OMOP domain: Condition (a), Drug (b), and Measurement (c). The pie charts illustrate how sparse the EHR data is for the MS patients. Note that Condition and Drug OMOP codes are empty for most MS patients. The violin plots show the distribution of log2 fold change (MS compared to Non-MS) values of the OMOP codes within each of these groups. The greatest fold change values are primarily in the groups of OMOP codes with only 1-2 MS patients. The most common MS symptoms (violin plot a) display negative log2 fold changes, demonstrating that they are common in the control population as well.

Supplementary Figure 3. Lab tests drive information to MS biology nodes. A commonly ordered lab test (aspartate aminotransferase) can increase the importance of the node *Myelin Sheath Adaxonal Region* by sending information through *L-Aspartic-Acid*, to *Diseases*, then *Genes* that are connected to *Myelin Sheath Adaxonal Region* (*MAG* and *PTEN*)

Supplementary Table 1. OMOP concepts that drive “all-OMOP” classifiers.

Supplementary Table 2. OMOP to SEP mapping coverage.

Supplementary Table 3. Biological nodes in SPOKE that drive the SPOKEsig classifiers and are higher ranked in the non-MS population.

Supplementary Table 4. Number of patients used in each analysis.

Supplementary Table 5. Comparison of SPOKEsig and SEP classifiers.