Journal of the American Medical Informatics Association, 2026, **33(1)**, 167–181
https://doi.org/10.1093/jamia/ocaf012
Advance access publication 31 January 2025
Research and Applications

Research and Applications

myAURA: a personalized health library for epilepsy management via knowledge graph sparsification and visualization

Rion Brattig Correia, PhD[1], Jordan C. Rozum, PhD[1], Leonard Cross, BS[2], Jack Felag, MS[1], Michael Gallant, MS[2], Ziqi Guo, MS[1], Bruce W. Herr II, BS[2], Aehong Min, PhD[3], Jon Sanchez-Valle, PhD[4], Deborah Stungis Rocha, MS[1], Alfonso Valencia, PhD[4], Xuan Wang, MS[2], Katy Börner, PhD[2], Wendy Miller, PhD, RN[5], Luis M. Rocha, PhD*[,1,6]

[1]School of Systems Science and Industrial Engineering, Binghamton University, Binghamton, NY 13902-6000, United States, [2]Luddy School of Informatics, Computing & Engineering, Indiana University, Bloomington, IN 47408, United States, [3]Donald Bren School of Information & Computer Sciences, University of California, Irvine, CA 92697-3435, United States, [4]Life Sciences Department, Barcelona Supercomputing Center, 08034 Barcelona, Spain, [5]School of Nursing, Indiana University, Indianapolis, IN 46202, United States, [6]Universidade Católica Portuguesa, Católica Biomedical Research Centre, 1649-023 Lisboa, Portugal

*Corresponding author: Luis M. Rocha, PhD, School of Systems Science and Industrial Engineering, Binghamton University, PO Box 6000, Binghamton, NY 13902-6000, United States (rocha@binghamton.edu)

Abstract

Objectives: Report the development of the patient-centered *myAURA* application and suite of methods designed to aid epilepsy patients, caregivers, and clinicians in making decisions about self-management and care.

Materials and Methods: *myAURA* rests on an unprecedented collection of epilepsy-relevant heterogeneous data resources, such as biomedical databases, social media, and electronic health records (EHRs). We use a patient-centered biomedical dictionary to link the collected data in a multilayer knowledge graph (KG) computed with a generalizable, open-source methodology.

Results: Our approach is based on a novel network sparsification method that uses the metric backbone of weighted graphs to discover important edges for inference, recommendation, and visualization. We demonstrate by studying drug-drug interaction from EHRs, extracting epilepsy-focused digital cohorts from social media, and generating a multilayer KG visualization. We also present our patient-centered design and pilot-testing of *myAURA*, including its user interface.

Discussion: The ability to search and explore *myAURA*'s heterogeneous data sources in a single, sparsified, multilayer KG is highly useful for a range of epilepsy studies and stakeholder support.

Conclusion: Our stakeholder-driven, scalable approach to integrating traditional and nontraditional data sources enables both clinical discovery and data-powered patient self-management in epilepsy and can be generalized to other chronic conditions.

Key words: epilepsy; self-management; semantic web; systems analysis; social media; data visualization.

Background and significance

The chronic health disorder epilepsy affects more than 3.4 million people in America and 65 million worldwide.[1,2] People with epilepsy (PWE) are at risk for lower quality of life, social isolation, depression, anxiety, medication-related symptoms, and premature death.[1,2,3–10] Exacerbating these risks, PWE can wait up to 9 months to see a neurologist and much longer to see an epileptologist, so many PWE are treated by general practitioners while they wait.[3,11–14] Thus, to achieve desirable health outcomes, self-management by PWE and their caregivers (PWEC) becomes essential[3] and they seek information online. While much recent research has aimed to help patients retrieve health information online, the sheer abundance from heterogeneous data sources makes it difficult for PWEC to distinguish the best treatment options available or even the relevance of information to an individual case. They are challenged by a daunting array of options about treatments, drugs, drug interactions and side effects, diet, lifestyle, and stigma.

Indeed, any chronic health condition unfolds as a complex interplay among all these biological, psychological, and societal factors that change over time. A personal health library with integrated and individualized information retrieval has a clear role to play in improving health outcomes for PWE and anyone with a chronic health condition. Qualitative and quantitative studies,[7,15,16] including those pursued under our project,[17,18] show a clear need for visually engaging, easy-to-use online tools for 2 key purposes: first, to extract, classify, organize, and personalize

Received: August 1, 2024; **Revised:** December 6, 2024; **Editorial Decision:** January 8, 2025; **Accepted:** January 14, 2025

information; and second, to provide automated recommendations in support of evidence-based decisions about treatment and self-management.

Here, we present milestones of the ongoing *myAURA* interdisciplinary project that aims to address this problem directly with data- and network-science methods to integrate multiple resources into a personalized easy-to-use web service for PWEC. To design this service according to their needs, our interdisciplinary team of experts in biomedical informatics, text and social media mining, visualization, user interface design, and epilepsy self-management works with patients, caregivers, and their advocates. We also leverage an exclusive use agreement with the Epilepsy Foundation of America (EFA) both to obtain PWEC data from their website, discussion groups, and social media presence, and to recruit PWEC for our user study group. All of this goes into computing a large-scale epilepsy *knowledge graph* (KG) that comprises a set of networks of associated data from heterogeneous data sources relevant to PWEC.

We compute the metric backbone, a network sparsification method based on removing edges that are redundant for shortest path computation,[19] of the epilepsy KG and discuss it as a powerful way to infer, identify, visualize, and recommend personalized, relevant information for PWEC. We also summarize our patient-centered methodology for designing a *myAURA* application. Per stakeholder needs and human-centered design specifications, when fully deployed, *myAURA* will integrate practical, location- and patient-specific health care information with targeted scientific literature, biomedical databases, social media platforms, and epilepsy-related websites with information about specialists, clinical trials, medications, community resources, and chat rooms.

Objectives

The innovative data- and network-science methods that underpin *myAURA* drive 3 research aims:

1) Produce a multilayer epilepsy KG ("Building the *myAURA* Epilepsy KG") of relevant terminology ("Biomedical Dictionaries and Sentiment Analysis") by federating heterogeneous sources of large-scale data ("Data Federation and Processing") and exemplify its value in the study of drug-drug interaction (DDI; "Studying DDIs Using KGs").
2) Develop recommendation and visualization algorithms by automatically extracting the epilepsy KG's metric backbone ("Analysis of *myAURA*'s KG Backbones" "Maps of *myAURA*'s KG").
3) Design and pilot test *myAURA* using focus group studies that survey PWEC regarding their desired *myAURA* content and its format ("User-Centered Design and Pilot Testing of *myAURA* Through Focus Groups").

Our immediate goal is to produce and visualize an epilepsy KG representation of heterogeneous resources in support of a user-friendly web service to facilitate PWEC self-management. We report on the interface design based on PWEC focus group input and the design requirements for other similar applications. Our long-term goal is to generate a personal health library for PWEC and in so doing create a suite of methods that can be generalized to support self-management of other chronic diseases.

Data and methods

Developing useful patient-centered tools requires integrating often disparate information resources and validating results by engaging end users. To ensure that *myAURA* meets the needs of PWEC, we federate on their behalf broad-ranging relevant data that we consider in 2 main groups, detailed below. We process the data from these resources to produce various large-scale knowledge networks[20,21] that are amenable to analysis with the powerful tools of network science[22–25] and machine learning.[26,27] We also engage with PWEC during development. The overall architecture is depicted in Figure 1.

Data federation and processing

Social media and community websites

We have previously demonstrated the utility of social media data in the study of epilepsy and other biomedical problems[10,28–33] and here include digital cohorts from Instagram, X (Twitter), Reddit, Facebook, YouTube comments, and the EFA website forums and Facebook discussion wall. For each, the epilepsy-specific digital cohort information refers to complete user timelines that contain all time-stamped posts of users who posted at least once about a drug used to treat epilepsy.

Instagram currently has more than 1.2 billion active users monthly. The current study uses a dataset collected between October 2010 and January 2016 via its API.[29] This epilepsy-specific digital cohort contains 9890 complete user timelines containing 8 496 124 posts.

From *X*, using the historical gardenhose and the OSoMe data and tool set,[34] we collected a random sample of 700K user timelines. Of these, the epilepsy-specific digital cohort contains 5958 complete timelines containing 14 152 929 posts.

Reddit is a user-moderated forum organized into over 100K subforums called subreddits that are devoted to specialized topics. The r/Epilepsy subreddit, devoted to PWEC, has been active since August 2010 and had more than 18K unique users who posted more than 277 367 times. (Though fewer, these posts are typically much longer than posts on Instagram or X.) For a more direct comparison with *Instagram* and *X*, we identified the epilepsy-specific digital subcohort of 6301 users. Their timelines contain a total of 219 459 posts that, unless otherwise noted, compose our epilepsy digital cohort for Reddit.

YouTube, used by an estimated 81% of Americans in 2021, is the most popular social media platform in the United States. Via its API, from a population of more than 330 347 users who engaged with epilepsy-related content over an 18-year period, we collected an epilepsy-specific digital cohort of more than 2035 users.

Overall, we collected over 48K complete user timelines, with over 23M posts for our epilepsy-specific digital cohort. With support from the EFA and via a specially developed application, we also collected from *Facebook* a small cohort of entire timelines of 12 victims of Sudden Unexpected Death in Epilepsy (SUDEP).

In addition, we have access to *the EFA website* (epilepsy.com), with more than 1M unique users per month, and its highly used message boards, chat rooms, comment threads, and the MyEpilepsyDiary (which allows users of the EFA website to track medications, seizures, triggers, side effects,

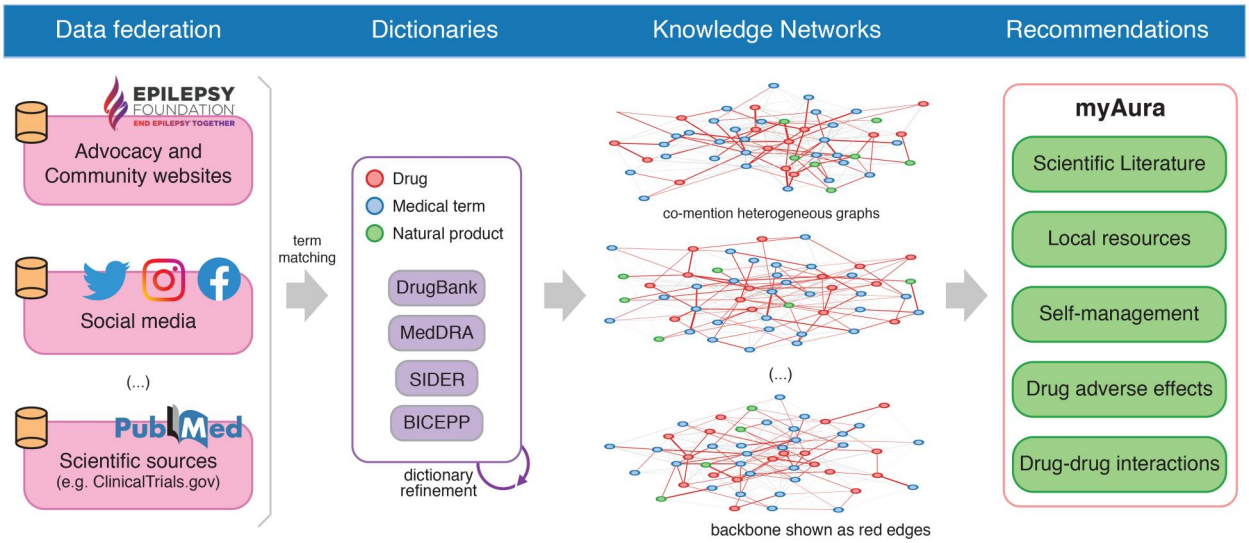

Figure 1. Diagram of the overall *myAURA* project: federated heterogeneous data; biomedical dictionary built from various scientific resources; the constructed multilayer epilepsy knowledge graph with its computed backbone; and the *myAURA* application features (local resources include, eg, epilepsy centers and clinical trials).

and symptoms) via an exclusive use agreement. The social activity on the site is akin to that on social media[28] with the added benefit to our research that users are focused on the target PWEC community and their activities and health considerations. Collected data are from 2004 to 2016, and it includes timelines of 22 938 active users with a total of 111 075 posts. The epilepsy-specific digital cohort (same inclusion criterion as for other social media) contains 8488 user timelines containing 78 948 posts.

Biomedical and patient data

In addition to social media data, our federated database includes clinical, pharmacological, health, and scientific databases relevant to epilepsy. We use anonymized population-wide *electronic health record* (EHR) data extracted directly from the public health care systems of the cities of Blumenau (Brazil, pop. 330K) and Indianapolis (USA, pop. 865K), and the whole of Catalonia (Spain, pop. 7.5M). We collected drug administration, patient demographics, and disease diagnoses (ICD-10) for 133 047 Blumenau patients for 18 months (January 2014-June 2015), 5 555 924 Catalonia patients for 11 years (January 2008-December 2018), and 264 607 Indianapolis patients for 2 years (January 2017-December 2018).[35,36] We addressed language differences by manually resolving drug compounds to their English names with the assistance of local medical experts and lists of drug synonyms obtained from *DrugBank*.[37] Medications with multiple drug compounds were split into their constituent drugs. Administered substances not found in DrugBank were discarded. For instance, in the Catalonia dataset, drugs are identified by their *Anatomical Therapeutic Chemical* (ATC) classification, which we mapped to DrugBank IDs using the finest level of ATC detail: chemical substance. The precise multilingual resolution pipeline is detailed in literature.[35,36] We curate these EHR data to compute KGs used to uncover DDI and adverse drug reactions (ADR) by risk level (major, medium, and minor), gender, and age. Analysis of these graphs, discussed below ("Studying DDIs Using KGs"), has already revealed important sex and age biases in all 3 populations.[35,36] Including these drug and symptom KGs in the *myAURA* data federation allows us to focus on epilepsy-relevant DDI and ADR, as well as epilepsy-specific biases. Moreover, these graphs enable future comparisons and analyses of DDI, ADR, symptoms, and temporal comorbidity trajectories in *myAURA*'s user population with those observed in independent patient populations. This will facilitate issuing medication and symptom warnings to *myAURA* users and PWEC at large.[35]

PubMed is a service of the National Library of Medicine, a "free resource supporting the search and retrieval of biomedical and life sciences literature with the aim of improving health—both globally and personally." It includes over 35 million citations dating back to the 1860s. We process updated local copies of the entire PubMed/MEDLINE database (28M citations) and use them in the *myAURA* KG to enable the recommendation of relevant scientific literature (eg, abstracts, MeSH terms, and references related to medications) to PWEC.

ClinicalTrials.gov is a central registration site for both publicly and privately funded clinical trials operated by the National Institutes of Health that has been available to the public since 2000.[38] The full dataset is available online. We ingest a local copy of the data into the Scholarly Database at Indiana University and integrate it into *myAURA*'s federated database tagged with its dictionary for KG construction. As explained below ("Biomedical Dictionaries and Sentiment Analysis"), the dictionary construction requires processing several resources to tag and link relevant pharmacology and symptom information.

Via our PWEC user focus groups ("User-Centered Design and Pilot Testing of *myAURA* Through Focus Groups"), we identify and ingest other resources deemed most useful to patients, such as the American Epilepsy Society's *Find a Doctor* Database,[39] which contains geographic locations of all epileptologists in the United States. Furthermore, the design of *myAURA*'s user interface ("User-Centered Design and Pilot Testing of *myAURA* Through Focus Groups") includes local transportation information, integrating services such as taxi, Lyft, UBER, and other public transportation via their APIs or Google Maps.

Biomedical dictionaries and sentiment analysis

Constructing specialized dictionaries from which to automatically tag text of potential relevance is a key aspect of resource federation, and we do so here to define the nodes required to build the epilepsy KG. We do as others have done in studying depression using *X*,[40,41] first including terms obtained from clinicians and extracted from epilepsy-patient social media.[29] We then add dictionaries carefully curated by pharmacology and biomedical informatics experts with over 170K standardized terms from Food and Drug Administration drug labels, DrugBank,[37] SIDER,[42] BICEPP,[43] FAERS,[44] *MedWatch*,[45] *Drugs.com*, and a standardized medical terminology dictionary built from clinical notes and MedDRA.[46] Parent terms and synonyms are resolved in a hierarchical manner (eg, Prozac resolves to fluoxetine, and cold to nasopharyngitis). The resulting dictionary of clinical terminology has already been used to produce a publicly available annotated corpora of PubMed articles and text mining pipelines to extract experimental evidence of DDI.[29,47–50]

Clinical terminology, however, is not tailored to social media language so it can bias biomedical inference pipelines. In our context, it must be refined, which we did via human-centered curation in the social media context.[31] We use the resulting *myAURA* dictionary to tag the relevant text fields from the federated resources above. We then index them in a data warehousing system to easily link relevant concepts to text units, users, and all data fields. The tagged terms become nodes in the epilepsy KG described below ("Building the *myAURA* Epilepsy KG").

In parallel, we use several dictionary-based sentiment analysis tools such as ANEW,[51] VADER,[52] and LIWC[53] to tag each post with a mood state along sentiment dimensions including valence (happy/sad), arousal (calm/excited), and dominance (in-control/dominated).[32] This allows us to estimate individual and collective psychological mood states of the epilepsy digital cohorts and affords various types of health-related discoveries.[28] For instance, we studied the SUDEP cohort mentioned above and showed that certain sentiment measures, such as increased or altered verbosity, may be predictive of this serious outcome, an important result for stakeholders.[10]

Building the *myAURA* epilepsy KG

Given that the textual items of the federated data resources are tagged with dictionary terms, it is straightforward to build weighted graphs (networks), where edges denote a co-occurrence proximity measure (eg, of drugs and medical terms on social media posts or in EHRs) or its inverse, distance. Given the set X of all terms, we first compute a symmetric co-occurrence matrix, $R_w(X)$, whose entries r_{xy} denote the number of textual units w where terms x and y co-occur.[19,54] Unit w may denote a PubMed abstract,[55] a user timeline-window on Instagram,[28,29] or an EHR prescription period,[35,36] for example. The diagonal entries of this matrix, r_{xx}, denote the total number of times term x was mentioned in a unit of analysis with any other term in the dictionary X: $r_{xx} = \sum_{y \in X: y \neq x} r_{xy}$. To measure a normalized strength of association among the X terms, we compute a *proximity graph* $P(X)$ whose edge weights are given by the weighted Jaccard similarity[56,57] (other measures are possible[54,58–61])

$$p_{xy} = \frac{r_{xy}}{r_{xx} + r_{yy} - r_{xy}}, \quad (1)$$

where $p_{xy} \in [0, 1]$ denotes a *proximity* between 2 terms x and y. When the terms never cooccur in textual units w, $p_{xy} = 0$; when they always co-occur, $p_{xy} = 1$; and $p_{xx} = 1$ (reflexivity).

The proximity KGs are powerfully simple data representations of the relationships among different entities. Also, computing KGs is scalable because it depends only on pairwise comparison of vectors for each pair (x, y).[54,62] Though more complex measures of word embedding, including those based on neural networks[63] (rather than the normalized Jaccard co-occurrence matrix P), are possible,[64] our approach guarantees maximum explainability of the uncovered pathways that associate terms in inference, recommendation, and visualization ("Results"). We can trace edge weights back to the unique textual units w where a given term co-occurrence is observed. The ability to unequivocally link a given inference to its evidence (eg, in the interactive exploration of our KG maps in "Maps of *myAURA*'s KG") is most important in biomedical applications.[65] Indeed, we have used this word embedding approach to build competitive recommender algorithms,[54,66–68] biomedical text mining pipelines,[55,69–71] scientific maps,[72–76] automatic fact-checking,[58] and network inference in biomedicine.[19,29,35,36,77]

The full *myAURA* epilepsy KG, $\varepsilon = \{P^s(X)\}$, is the set of proximity graphs derived for each federated data source s. Because each constituent graph, P^s, is defined on the same node-set of dictionary terms $x \in X$, ε is a multilayer graph where term associations for each data source are represented separately, with interlayer edges connecting the same dictionary terms on each layer (ie, a multiplex as represented in Figure 1).

Source code, software, and data sharing policy

We developed *standardized open-source code* for efficient data ingestion, preprocessing, dictionary term matching, construction, storage, and joining of networks from different data sources into *PostgreSQL*, and for computing the metric backbone ("The Metric Backbone for KG Sparsification"). It provides a unique, fast, and streamlined process that reproduces the complete KG construction pipeline for each independent data source that gets updated. This source code and an open-source API for accessing an SQL database with shareable resources is publicly available through github (at github.com/cns-iu/myaura for KG construction and github.com/CASCI-lab/ for the backbone extraction pipeline), the most widely used repository of public software, with appropriate licenses that allow stakeholders and other researchers to reuse and build upon our source code. This is in keeping with the Findability, Accessibility, Interoperability, and Reusability[78] principles that are foundational to *myAURA*'s federated data.

Given the biomedical and health context, we must also address technical barriers and the ethical and legal considerations of human subject data and protected health information. Importantly, we must also safeguard our focus group participants ("User-Centered Design and Pilot Testing of *myAURA* Through Focus Groups"). These factors are considered at every step of our NIH-sponsored project via approved and detailed Data and Safety Monitoring and Protection of Human Subjects plans. When the data policy of specific resources, or ethical and legal considerations, prevent

redistribution, we release processing tools without raw data via github. These provide analytical support for the computation of recommendations, KGs, personalization, and visualization features. In this way, the data pipelines can be validated, and the analyses can be reproduced on alternate data sets via reusable software. This has been especially important in multi-institutional collaborations such as the study of DDI from EHRs ("Studying DDIs Using KGs").

Social media data are, in general, bound by changing sharing and reusability policies, which are distinct by platform. Our digital cohorts ("Social Media and Community Websites"), however, are stored in a secure SQL. From there, resources within the respective platforms can be pointed to and accessed based on their own policies. *Instagram*, for instance, does not prohibit the sharing of post URLs and their metadata, though we do need to respect their user copyrights. Therefore, instead of sharing Instagram media (photos and videos) directly, we share the URLs where this information is accessible. In contrast, *X*'s terms of service explicitly prohibit the sharing of raw tweets and metadata. In this case, we publicly share only aggregated and graph data via our API and the *myAURA* network visualization tool ("Maps of *myAURA*'s KG").

Results

Studying DDIs using KGs

We exemplify the utility of the *myAURA* KG with the networks obtained from EHR data. We conducted a large longitudinal study of the prevalence of known DDI in the Blumenau, Brazil dataset. Even after correcting for multiple factors, we found women and older patients were at significantly higher risk of DDI exposure than their polypharmacy regimens would suggest.[36] Then we showed the generalizability of our pipeline using additional primary care data from the distinct locations of Indiana, USA, and Catalonia, Spain. We found very similar sex and age biases in the prevalence of known DDI in both locations, albeit sometimes involving distinct drugs[35]; several drugs used in epilepsy were implicated (Figure 2). Moreover, our analysis revealed explainable and actionable interventions that easily reduce both biases and the burden of DDI (eg, replacing Omeprazole with another proton-pump inhibitor). To facilitate the reusability of our data and KGs, with a wider consortium, we built the DDInteract web tool[35,79] to enable third-party analysis of the DDI KGs from all the *myAURA* EHR datasets (Figure 2). In addition to integrating all the EHR-derived information about DDI, the tool allows contextual inference per population, age, gender, and severity and significance of interaction. For example, carbamazepine interacts with ethinylestradiol, possibly reducing the latter's blood levels and efficacy; however, this DDI is only relevant in the younger female population. DDInteract allows users to extract KGs for specific age ranges to observe how the DDI phenomenon changes contextually. It is important to emphasize that this comparative study and reusability demonstration required expert-driven multilingual resolution of drug names used in the various EHRs ("Biomedical and Patient Data"). While such a process is not trivial, with access to local medical experts and available drug and chemical classifications, we show that it is feasible to reconcile multilingual patient data.

In addition, we built text classifiers to identify PubMed abstracts (and sentences) with direct experimental evidence of DDI because our focus group user studies ("User-Centered Design and Pilot Testing of *myAURA* Through Focus Groups") revealed that detailed pharmacological information is of particular importance to epilepsy patients. We trained classical and large language models like BioBERT and ChatGPT on our human annotated DDI PubMed corpora[50] and on human-annotated Instagram posts,[31] and they performed very well.[48,80] The recommendation of relevant experimental DDI evidence supports functionalities in the *myAURA* app design ("User-Centered Design and Pilot Testing of *myAURA* Through Focus Groups"). When users click on nodes that represent drugs in the KG visualizations, they may, for example, obtain PubMed articles with experimental evidence of DDI for those drugs ("Maps of *myAURA*'s KG").

The metric backbone for KG sparsification

KGs are often dense with many edges that are not relevant for inference and, furthermore, impair visualization and slow down computation. Therefore, we developed a sparsification method to facilitate analysis and visualization of *myAURA*'s epilepsy KG and other network data-based informatics problems.

Many network inference methods depend on computing shortest paths on *distance graphs* $D(X)$, whose edge weights, $d_{xy} \in [0, +\infty]$, denote an antireflexive distance between nodes x and y ($d_{xx} = 0$). These are easily obtained from the proximity graphs of ε ("Building the *myAURA* Epilepsy KG") via

$$d_{xy} = \frac{1}{p_{xy}} - 1, \quad \forall x, y \in X. \tag{2}$$

The resulting distance weights are inversely proportional to the strength of association between terms; thus, they convey a measure of distance necessary to compute path length. It is typically symmetric, but the approach also applies to the asymmetric associations (ie, directed graphs).[81]

Shortest paths allow us to infer the strength (or cost) of indirect association: If x is connected to z with a finite distance, and y is similarly connected to z, the length of the shortest indirect path quantifies how close x is to y via z. This type of inference is ubiquitous in network problems.[22–25] We have shown that distance graphs obtained from real-world data are typically not metric, but rather *semimetric*,[54,66] meaning the *triangle inequality* ($d_{xy} \leq d_{xz} + d_{zy}$) is not observed for every edge of $D(X)$.[82] That is, the shortest distance between at least 2 nodes in the graph is not the direct edge, but rather an indirect path via other nodes.

Computing shortest paths of a distance graph, where path length is the sum of constituent edge (distance) weights ($d_{xy} = d_{xz} + d_{zy}$), for example via Dijkstra's algorithm,[83] yields its *metric closure* $D^C(X)$, which obeys the triangle inequality at every edge.[54] If an edge in the original graph is semimetric, its weight gets replaced by the length of the shortest indirect path between the nodes it connects. If an edge weight d_{xy} of $D(X)$ does not change in the metric closure $D^C(X)$, it is *metric* because it obeys the triangle inequality—there is no indirect path shorter than the direct edge between x and y—while those that change, are semimetric.

Significantly, there is a *metric backbone* subgraph $D_b(X)$[19] of the original graph $D(X)$ that is invariant under the metric closure and is sufficient to compute all shortest paths:

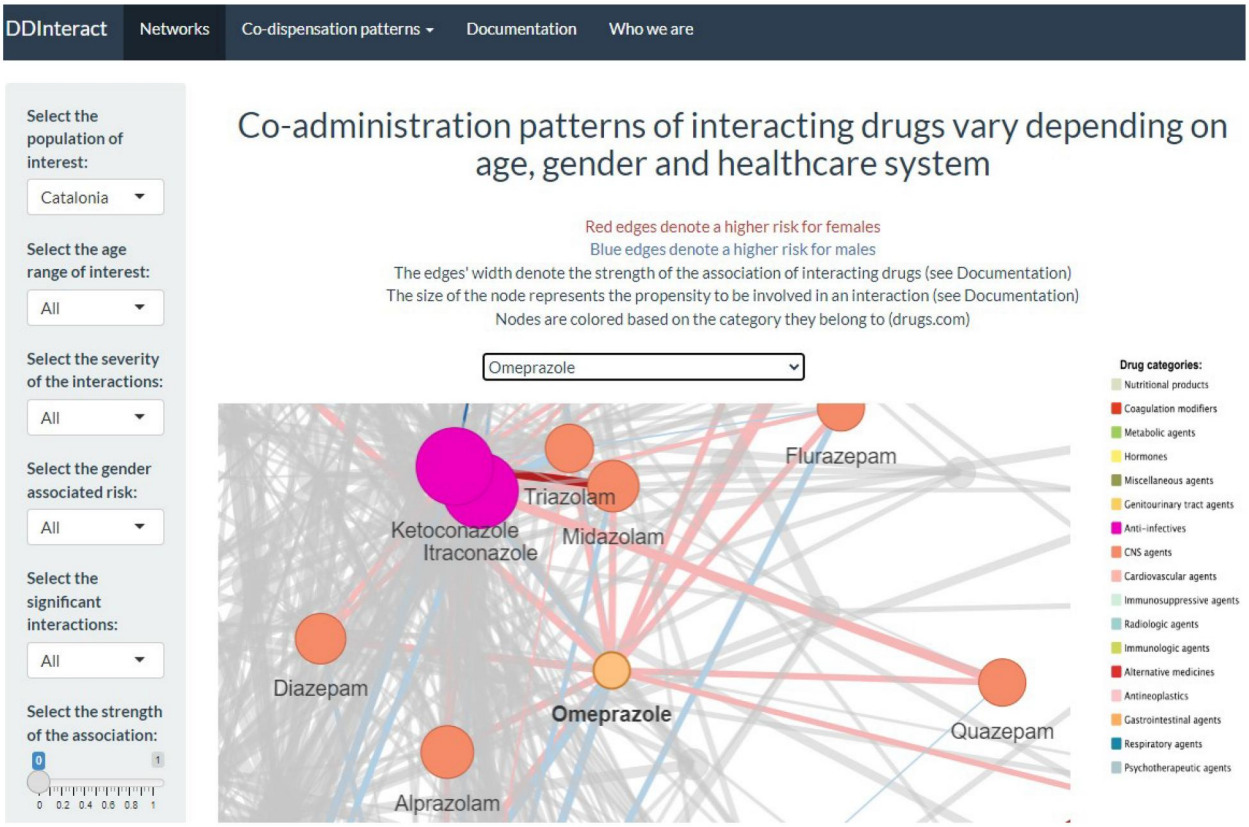

Figure 2. A DDInteract tool[35,79] screenshot displaying the Catalonia DDI (proximity) network extracted from EHRs, highlighting the most frequent interactions associated with Omeprazole.

$D_b^C(X) \equiv D^C(X)$. The edge weights of the metric backbone graph are given by

$$b_{xy} = \begin{cases} d_{xy}, & \text{if } d_{xy} = d_{xy}^C, \\ +\infty, & \text{if } d_{xy} > d_{xy}^C, \end{cases} \tag{3}$$

where $b_{xy} = +\infty$ means there is no direct edge between x and y.

Edges not on the backbone are superfluous in the computation of shortest paths (and in all network measures derived from shortest paths). Importantly, the metric backbone is an algebraically principled network sparsification method with unique features in that it (1) preserves all connectivity and shortest path distribution, (2) does not alter edge weights or delete nodes, (3) is exact, not sampled or estimated, and (4) requires no parameters or null model estimation.[19] Furthermore, it outperforms available state-of-the-art network sparsification methods in that it (5) preserves the community structure (modularity) of the original graph,[84] and (6) recovers most of the original (macro and micro) spreading dynamics, reveals the most important (information or disease) transmission pathways, and results in greater reduction without breaking apart the original network.[19,77,85]

The size of the backbone subgraph in relation to the size of the original graph defines the amount of *redundancy* in the network. All layers of *myAURA*'s KG have a small backbone (large amount of redundancy) as seen in Table 1 and Figure 3. This is coherent with what is observed in networks across biological, technological, and social domains, which typically possess very small metric backbones.[19] For instance, the metric backbone of a protein interaction network of more than 11K human genes involved in spermatogenesis comprises ≈ 10% of the original edges.[77] The 90% of edges not on the backbone were obtained from experimental evidence, but they are redundant for shortest paths and likely less important for regulatory pathways. This realization led to our discovery of new genes involved in male infertility.[77] Similarly, the backbones of social contact networks important for epidemic spread are within 5%-20%[19,84]; those of the human brain connectome and functional and multiomic gene co-expression networks are typically 5%-11%[19] and have distinguishing network features that enable effective classification between healthy and diseased human cohorts in Alzheimer's disease, autism, depression, and psychotic disorder.[86–88]

Table 1. Network parameters of layers of the *myAURA* KG for various data sources.

KG network	Nodes	Edges	% metric
PubMed[a]	8891	590 781	18.59
Clinical trials	1275	31 371	53.75
Instagram	1686	25 235	15.1
X (Twitter)	1022	5082	37.0
r/Epilepsy (Reddit)	1270	17 558	17.0
EFA	1529	33 795	15.7

Columns show the numbers of nodes, edges, and size of metric backbone as the proportion of edges kept.
[a] Only epilepsy-related publications.

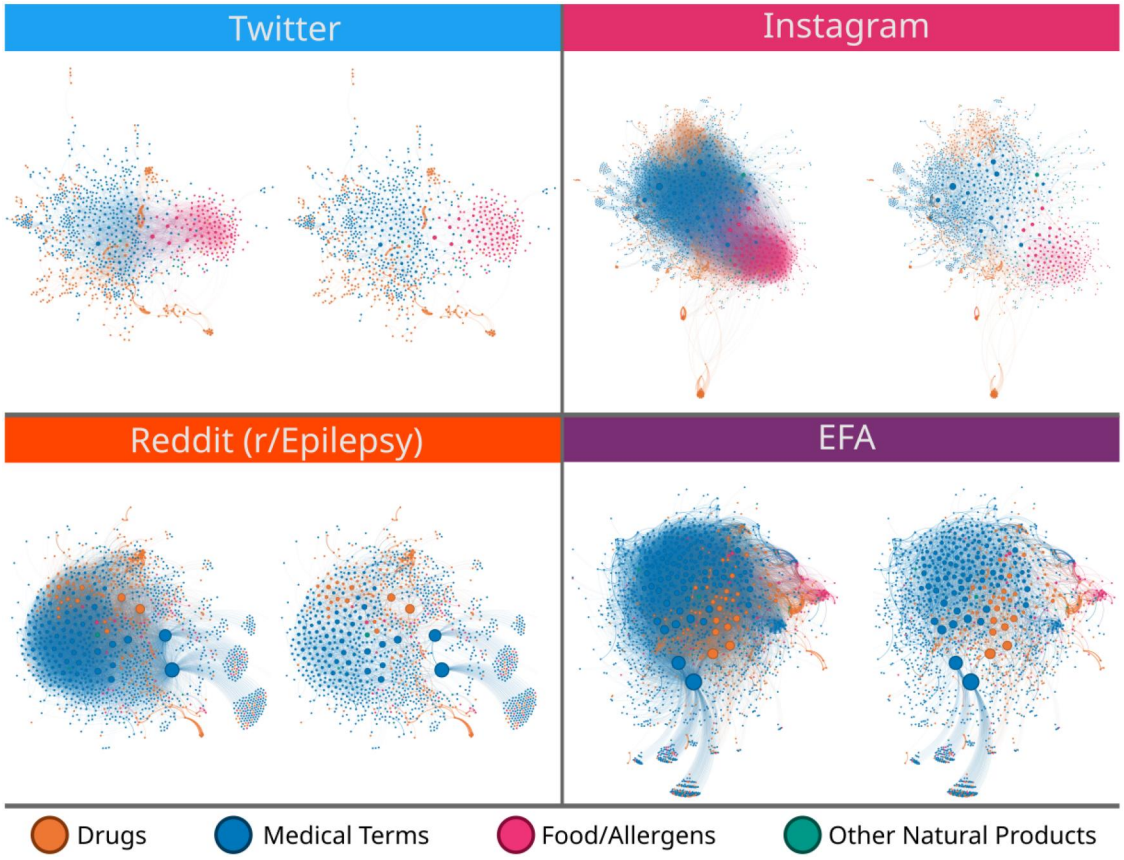

Figure 3. Original distance graph (left) and its metric backbone subgraph (right) for each of the social media layers of *myAURA*'s multilayer KG. Nodes are sized according to (unweighted) degree in the original network. Node positions are determined by the *Force2Atlas* method applied to the original networks. Only the largest component of each network is shown.

These observations show that the metric backbone is more than a mathematical construct. It has a *phenotype* in that its measurement in many biomedical and social complexity problems reveals important functional characteristics, such as community structure, information spreading dynamics, and the most important (central) network nodes, edges, and pathways for inference.[19,84,85] Additionally, because the backbones of large networks are typically very small, this natural sparsification provides substantial memory and computational parsimony in network storage and analysis.[62] Removing edges that are redundant for shortest paths clearly yields a powerful sparsification methodology that facilitates analysis and visualization of KGs,[19] which we illustrate next ("Analysis of *myAURA*'s KG Backbones"). We provide an open-source Python package for metric backbone extraction and analysis, *DistanceClosure*,[89] which is compatible with *NetworkX*, for interoperability with common graph formats (eg, *GraphML*, *GML*).

Analysis of *myAURA*'s KG backbones

Social media sites are useful in studying the interplay between human behavior and medical treatment in chronic diseases such as epilepsy,[10,28] but vary in the generality of their discourse. While *X* and *Instagram* engage a wide range of topics simultaneously, *Reddit* subgroups and the EFA discussion forums focus more on health-related discourse. Interestingly, sparsification of the derived KGs reveals these discourse differences. The metric backbones of the *myAURA* KGs from *Instagram*, *X*, *r/Epilepsy*, and EFA forums are similar in size: $\approx 16\%$ of original, except for X, which is 37% (see Table 1 and Figure 3). However, while harvested with the same criterion (at least one post mentioning an epilepsy drug), the proportion of users who contribute to backbones (at least one post containing a pair of dictionary terms represented by an edge on the backbone) is quite distinct across platforms. A much higher proportion of users contribute to the backbone in epilepsy-focused than in general-purpose social media: 65% and 71% on *Instagram* and *X* versus 95% and 93% on EFA forums and *r/Epilepsy*.[90] Thus, general-purpose platforms have a lot more users who contribute to redundant edges and not to any shortest path inference on the derived KGs.

We further observed that there is a clear discourse distinction between users who contribute to the backbone and those who do not. Using the human-annotated corpus of *Instagram* posts utilized to refine the *myAURA* dictionary (see "Biomedical Dictionaries and Sentiment Analysis"),[31] we observed that the false positive rate (dictionary terms used without medical relevance) is significantly higher for the set of users who do not contribute to the backbone (32%) than for those who do (14%). This difference is not a matter of engagement; false positive rates are similar for users who post a lot (13%) or little (18%). Thus, by extracting focused digital cohorts from general-purpose social media, such as *X* or *Instagram*, that are more like those on special-purpose forums of biomedical relevance, such as the EFA and

r/Epilepsy, the metric backbone sparsifcation of KGs can be used to increase personalization of social media data for a specific health problem.[90]

The sparsification of KGs into their metric backbone subgraph can also reveal key drug and medical term associations in the epilepsy patient discourse. For instance, all nodes shown in Figure 4A are directly associated with the term "Cannabis" because they co-occur in posts on r/Epilepsy, resulting in a confusing "hairball" graph. In contrast, the metric backbone of this graph, shown in Figure 4B, is strikingly simpler to study. Importantly, no reachability or shortest path information is lost in this sparsification. All terms on the original network remain reachable from "Cannabis" via an indirect path on the full backbone KG in Figure 3 with the exact same shortest distance (all shortest paths are preserved in the backbone, "The Metric Backbone for KG Sparsification"). However, most terms in Figure 4A are no longer directly connected to "Cannabis" in Figure 4B. The nodes that remain directly connected to "Cannabis" in the backbone have a transitive relationship to it in that the direct distance is shorter than or equal to any indirect path. Thus, their association is strongly and directly observed in user posts, not because of co-occurrences with other indirect terms. Similarly, many edges between nodes disappear in the "Cannabis" ego-graph backbone because their relationship via this target term is stronger than any direct co-occurrence measurement. From the perspective of shortest paths, the dictionary terms in the "Cannabis" ego-graph backbone are those most relevant to understanding how the *r/Epilepsy* digital cohort discusses this term directly.

Looking at another data source, Figure 4C shows a subgraph of the EFA KG backbone that includes all the nodes directly associated with two epilepsy treatment drugs, *Levetiracetam* and *Carbamazepine* (larger purple nodes), that are frequently prescribed together in drug-resistant epilepsy. In the middle of the graph are terms near both drugs, including additional drugs often co-prescribed with these medications, and medical terms related to side effects. Some of the terms, such as *mood swings*, *aggression*, *depression*, and *crying*, are moderate to severe side effects more associated with *Levetiracetam* (often main reasons patients switch from this medication) and they appear in close proximity to it in the graph. *Nightmares* are a common side effect of both drugs, but typically worse with *Carbamazepine*, coherent with the data in the subgraph. Indeed, there is no direct backbone connection between *Nightmares* and *Levetiracetam*. Also interesting, *Carbamazepine* is not safe to take during *Pregnancy* as it is associated with neural tube defects, but *Levetiracetam* is considered safer. Many women who plan to become pregnant switch medications during this time and the term appears between both drugs.

These examples highlight how the metric backbone can be leveraged to improve our understanding of how patients discuss drugs and drug side effects in various social mediums. The insights can be relevant to both patients and clinicians. Indeed, ego networks and other KGs are being used in researching various mental health and biomedical problems.[91] The metric backbone sparsification facilitates and improves such analysis because the resulting subgraphs are easier to visually inspect, without any loss to the shortest path distribution. All inferences based on distance associations among all terms revealed by the data are preserved.

Maps of *myAURA*'s KG

The full *myAURA* KG, ϵ, includes various networks extracted from distinct data sources and units of analysis (ie, EFA comments, tweets, or paper abstracts). For maximum explainability, *myAURA* enables tracing back the specific discourse that gave rise to an edge (ie, provides direct access to the context in which the terms were used). The impressive parsimony of the metric backbone makes its paths ideal *lines of argumentation* for explaining why certain inferences are made.

Thus, we developed a backbone-based *myAURA* KG visualization tool[92] using the *map4sci* visualization suite.[76,93] Using the *Zoomable Multi-Level Tree* algorithm,[94,95] it charts the knowledge embedded on backbone subgraphs onto a 2D plane by combining edges from each constituent network $\{P^s(X)\}$ according to a specific aggregation operation.[19] In the current implementation, p_{xy} values from each layer are averaged across data sources, but other aggregations are possible (eg, choosing the maximum p_{xy} or minimum d_{xy}; see eqns (1) and (2)) in all layers as we have done in the aggregation of multilayer protein-protein networks in another setting.[77] The result resembles a cartographic map with 3 graph layouts: *BatchTree*, which optimizes for scalability using `C++` and `OpenMP`, balancing between a compact layout and edge length preservation; *CG*, which optimizes compactness at the expense of preserving edge length; and *DELG*, which optimizes preserving edge length.

All variations are based on the metric backbone of KGs and use the same visual metaphor that displays semantic countries (clusters of strongly associated dictionary terms) that contain cities (terms) linked by roads defined by shortest paths connecting them (the most important associations for information transmission); see Figure 5A. We have shown with human subject studies that such map-like visualizations are as good as or better than standard node-edge representations of graphs in terms of task performance, and in memorization and recall of the underlying data.[96] Notice that semantic countries are mostly unaffected by sparsification because the metric backbone preserves community structure.[84]

As a user zooms in, edges down the hierarchy of importance (larger distances or lower proximity) are revealed as peripheral roads between lower importance dictionary term nodes. The tool provides easy dictionary-term searching in the map, for example, "Vagal nerve stimulator implantation" as depicted in Figure 5B. The online version of the map visualization tool also allows clicking on edges to retrieve information associated with the connected terms. However, due to privacy and access rights for each data source, the public version does not retrieve the actual data items where the terms are comentioned. Only our private, PHI-compliant research prototype is able to retrieve ranked data from all included resources after clicking on specific nodes and edges. The ability to search and explore *myAURA*'s heterogeneous data sources via a single combined map, as depicted in Figure 5D, is a useful feature of this visualization approach.

User-centered design and pilot testing of *myAURA* through focus groups

Our priority has been to understand and include the needs of PWEC in prototyping an application to support epilepsy self-management. We carried out a series of focus group interviews to understand how to deliver personalized

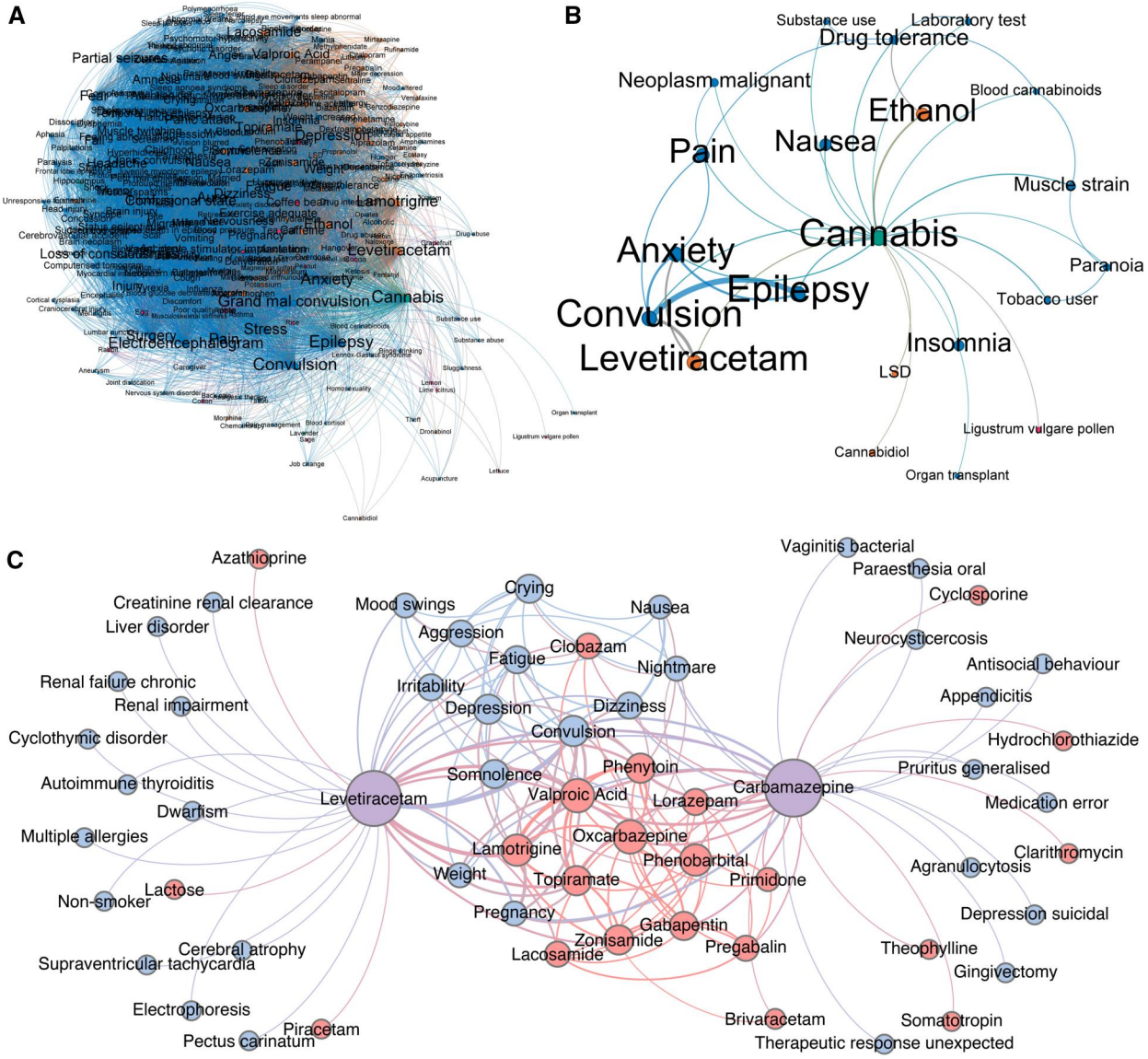

Figure 4. (A) "Cannabis" ego network using a Force2Atlas layout. (B) The metric backbone sparsification of A using a manual layout. Both are subgraphs of the full *Reddit* KG counterparts in Figure 3. (C) A subgraph of the EFA KG backbone in Figure 3, depicting terms directly connected to both *Levetiracetam* and *Carbamazepine*, 2 drugs commonly prescribed to treat epilepsy. Node colors denote dictionary term type: medical term (blue), drug (red), or queried term (purple).

recommendation and visualization of information from *myAURA*'s KG. The initial focus group had 12 PWEC participants that met for 4 sessions. They reported having difficulty finding the right information due to diverse symptoms among PWE. Participants also had trouble tracking and managing epilepsy-related information because it is distributed among multiple sources. Finally, they reported difficulty in sharing information with doctors and family members and in getting support during and after having seizures.[17,18] Participants were eager for an application like *myAURA* to be an epilepsy-specific all-in-one platform to track symptoms, seizures, available treatments, and other relevant factors, and to provide them and caregivers with a holistic image of their epilepsy status.[17,18] The ability to obtain personalized information (eg, the most effective treatments based on their individual symptoms and contexts) was also very important for them, as was the ability to share information easily with family members, friends, teachers, and health care providers.

A second focus group (a subset of the original group) tested our initial interactive mockup prototype containing the desired key features. We provided access at a virtual meeting and asked participants to perform a few tasks while the screen was shared with the researcher. Then they used the mockup freely for several minutes before sharing their experience in a short follow-up interview. Finally, they completed a survey designed to measure their perceptions and experiences with the mockup on a 7-point Likert scale.[97] Overall, the interview and survey results were positive. Higher scoring items were easiness to learn (6.38), feeling of control (6.19), and overall impression (6.17); relatively lower, though still positive scoring items were creativity (5.19), usefulness (5.34), and satisfaction (5.36).

Based on valuable participant feedback, we designed and implemented a final interactive mockup *myAURA* app (see Figure 6). It includes trackers (eg, food/water intake, sleep, menstrual cycles), modifications to the navigation of screens,

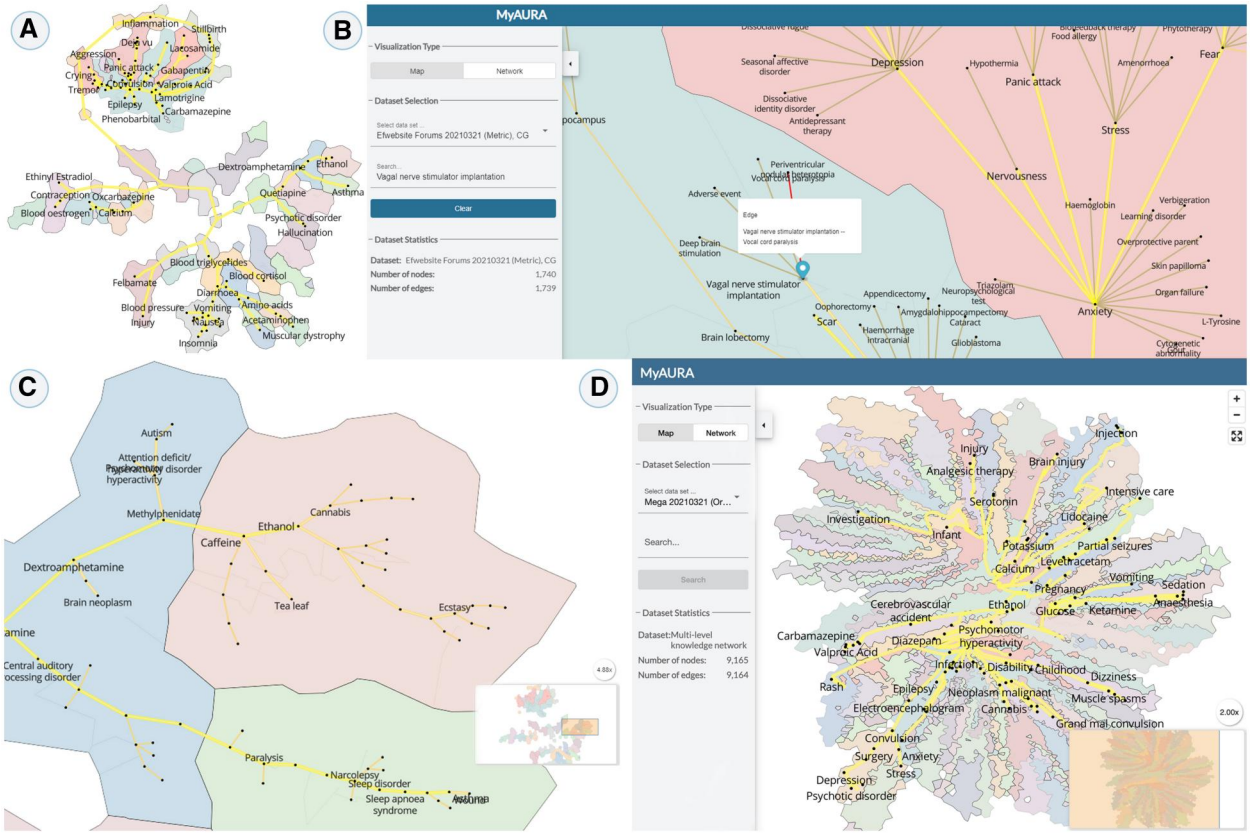

Figure 5. Map visualization tool for *myAURA* KG.[92] (A) Metric backbone of EFA forums KG at the top level. (B) Search and zoom in on term "Vagal Nerve Simulator Implantation," which is a "town" in "surgery country" (green), neighboring the "anxiety/depression country" (pink). (C) Zoom in (4.88×) on "Ethanol country" (pink), neighboring "dextroamphetamine country" (blue), and "asthma country" (green); the inset shows the top level map with the zoomed portion highlighted. (D) The combined ε multilayer KG, shown at 2× zoom level. This enables interactive exploration of terms indexing data items about clinical trials, drugs, and diseases extracted from multiple sources and their relations. The interface can be accessed at cns-iu.github.io/myaura.[92]

a dedicated media library where users may curate epilepsy-related information, an emergency/seizure response function, the ability to sync the platform with fitness trackers (eg, FitBit), and appointment/medication reminders. These items were deemed fundamental to providing meaningful information to PWEC. Then, to identify additional potential functions and to understand seizure management experience in diverse environments (eg, home, school, workplace, public transport), we conducted a third study with a newly updated mockup.[17] Our aim was to better understand the contexts, challenges, and coping strategies for seizure management devised by PWEC. We focused on understanding the social stigma experienced by PWE and proposed human-computer interaction design requirements to effectively deliver appropriate first aid information to bystanders during a seizure.[17]

Further, the 3 studies allowed us to complete a system-design framework to characterize challenges PWEC face in finding personalized just-in-time information, and tracking and sharing it with family, caregivers, and others. With this human-centered approach, we proposed a design framework to mitigate the challenges PWEC face and improve epilepsy information management and care coordination in *myAURA* or a similar future technology.[18]

We are committed to disseminating *myAURA* and our project outcomes to PWEC, clinicians, and researchers as widely as possible. However, direct communication with PWEC on social media is ethically problematic as it may result in unintended consequences for human subjects in uncontrolled settings. Therefore, in forthcoming work, we will significantly widen focus group studies to include epilepsy clinicians and researchers. They will guide and evaluate our data-powered approach toward desirable translation to scientific and health care insights. Furthermore, members of the EFA Professional Advisory Board, composed of neurologists, epileptologists, nurse practitioners, social workers, and psychologists, are enthusiastic and committed to supporting development, testing, and dissemination of not only the application, but also our overall analyses. They have agreed to disseminate the project's research findings and suite of tools to stakeholders via their website and social media channels, including *Facebook* and *YouTube*. The EFA website, epilepsy.com, has 1M+ unique users per month, making it the most widely used resource for PWEC. It provides the ideal mechanism, not only to recruit focus group participants but also to disseminate *myAURA*'s suite of tools and other analyses, empowering patient self-management as widely as possible.

Discussion

Our interdisciplinary efforts toward building *myAURA*, a personalized easy-to-use web service for PWEC, are ongoing. Most of our efforts so far have been in researching novel and explainable data and network science methods to design and implement the computational architecture of *myAURA* as a user-friendly web service to improve patient activation.

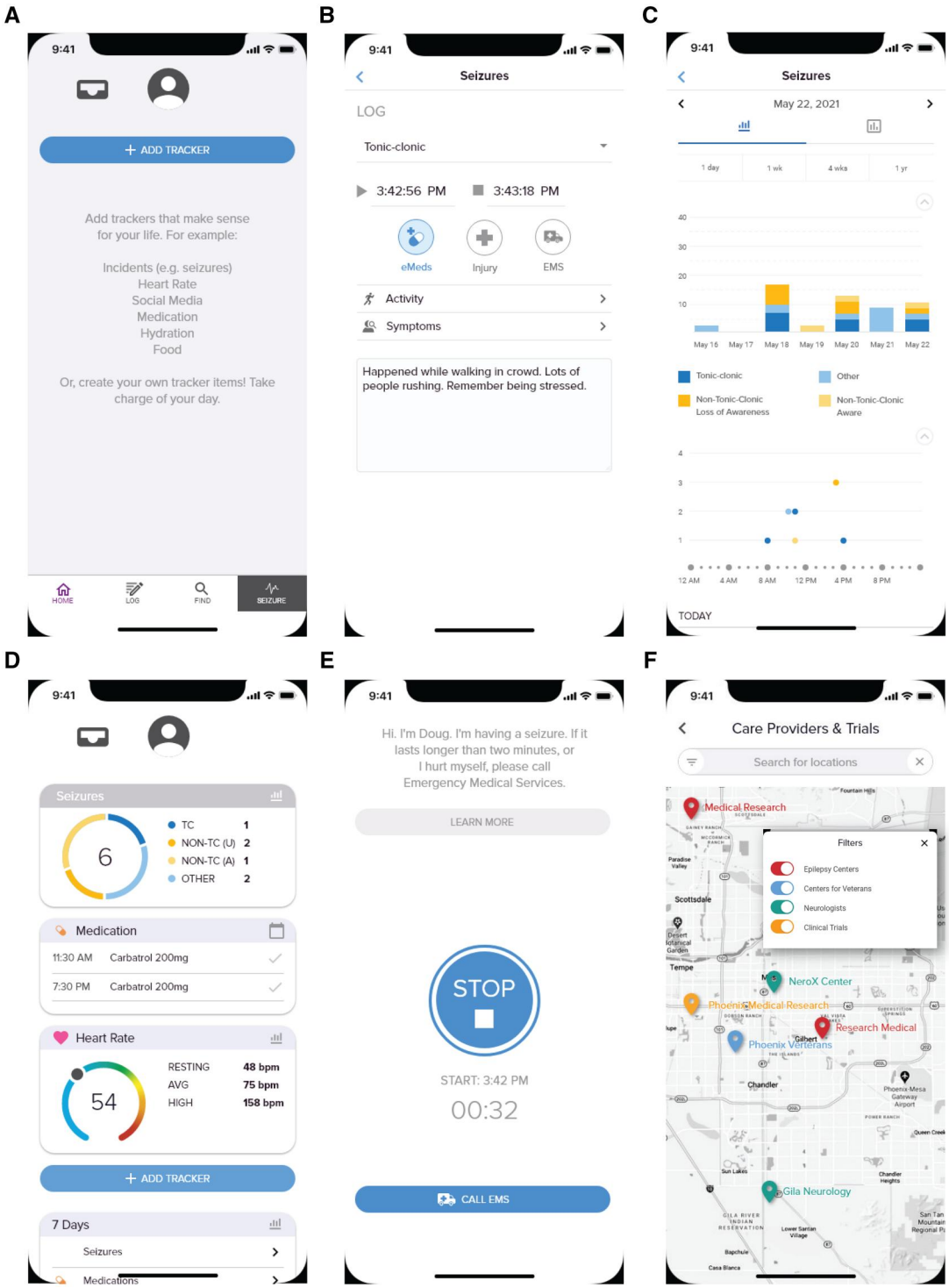

Figure 6. The latest user interface of the *myAURA* application. (A) A selection of items users can track (eg, seizures, medications, food). (B) Seizure-log data-input screen. (C) Data visualization for different types of seizure log. (D) User-customizable dashboard of important items, from seizure statistics to medication alerts and current heart rate. (E) Automatic message alert system that notifies emergency contacts in the event of a prolonged seizure. (F) Map search results interface, where epilepsy centers, neurologists, and clinical trials can be found and filtered.

We have translated this into several important novel developments discussed next.

Our approach rests on an unprecedented collection of large-scale heterogeneous data resources of relevance to study the specific biomedical and social complexity of epilepsy, in support of PWEC, including social media and community websites, EHR, and biomedical databases ("Data Federation and Processing"). To integrate so much data, we developed a generalizable methodology to compute a multilayer KG ("Building the *myAURA* Epilepsy KG"), based on the federation of the constituent heterogeneous data sources ("Data Federation and Processing") in separate layers linked via the terms of a patient-centered biomedical dictionary ("Biomedical Dictionaries and Sentiment Analysis").

We exemplify the power of this KG approach with DDI studies in EHRs,[35,36] the scientific literature,[50,80] and social media[28,29] ("Studying DDIs Using KGs").

To analyze the multilayer KG, we developed a network sparsification method (with corresponding open-source code) that allows us to extract the *metric backbone* of KGs, which removes edges redundant for shortest paths. It outperforms existing network sparsification methods ("The Metric Backbone for KG Sparsification", features [1-6]) and uncovers the most important edges and pathways for inference, recommendation, and visualization.[19,84,85] In addition, we show that metric backbones of KGs reveal how patients discuss disease factors and pharmacology on social media 4.3, which has led to another novel method to extract focused digital cohorts, whose discourse is more relevant to epilepsy, from general-purpose social media.[90]

The metric backbone is particularly amenable to simplifying the visualization of network data.[19] Applying this advantage, we developed geospatial map-like visualizations of sparsified KGs that enable the intuitive exploration of networks,[94,96] interactive search and extraction of relevant underlying data items, and merging of *myAURA*'s multilayer KG into a single map ("Maps of *myAURA*'s KG"). This integration provides the ability to search and explore *myAURA*'s heterogeneous data sources via a single sparsified and combined map in an easily consumable visual format useful for both epilepsy clinicians and PWEC. The optimal network reduction provided by our network sparsification approach is vital to *explaining* the inferences provided to *myAURA* users because it allows us to parsimoniously integrate, analyze, and simplify large amounts of multiscale data in real-time. Importantly, our approach enables direct retracing of all supporting evidence from heterogeneous data sources. This supports patients and clinicians in understanding multiscale factors involved in epilepsy. Indeed, the design of *myAURA* allows its users to query multiple different sources of information about topics such as medications, side effects, scientific literature, and clinical trials all on one platform. The resulting visualization displays relationships among these important topics acquired from a robust combination of data sources. Such large samples of social media and discussion forums related to epilepsy, clinical trials, or scientific literature would not typically be accessible to PWEC. Clinicians such as neurologists, epileptologists, nurse practitioners, physician assistants, and psychologists may also use *myAURA* to quickly and easily visualize knowledge about practice-relevant topics affecting PWE. These visualizations could be used quickly, even during a patient encounter, to guide assessments or treatment recommendations. In particular, once we have addressed the privacy and access rights for each data source in an updated working tool, visualization of a combination of data from social media and EHRs can reveal relationships among issues important to the PWE patient in association with their health records.

In forthcoming work, we will validate these sparsified visualizations and different methods of combining multilayer edges with both PWEC and epilepsy researchers, focusing on the value of deriving explainable inferences. The current multilayer KG constitutes a multimodal integration of many heterogeneous data resources, which are relevant to epilepsy and span multiple scales, from molecular and pharmacological evidence in the scientific literature and EHRs, to human behavior extracted from social media. Our methodology can extend in a straightforward manner to any data that can be labeled with our dictionary terms ("Biomedical Dictionaries and Sentiment Analysis"), such as imaging in EHRs, video, and audio. Tagging medical images is already feasible[98] and video sources, such as *YouTube*, often provide transcripts that make such integration possible in the future. Indeed, for a forthcoming version of *myAURA*, we are working on integrating a *GraphRAG*[99,100] chatbot to leverage our multilayer KG with the power of large language models. This will add a new user-friendly interface enabling patients and clinicians to interact with *myAURA*'s structured, multiscale, biomedical knowledge. Additionally, we will test Google's *NotebookLM* to produce engaging podcasts from textual documents such as the scientific articles recommended by *myAURA*, similar to those recently done for other health problems.[101] Both approaches may lead to improved self-management for users who are not familiar with or dislike reading scientific documents. In planned future focus groups, we will evaluate the putative benefits and address the risks of generating false or unexplained recommendations that are possible with such AI methods. We note that these kinds of risks do not occur with our current explainable approach.

We will continue our established approach of relying on stakeholder input at each step, whereby *myAURA*'s functionalities and interface were developed and pilot tested with patient-centered design principles based on focus groups studies ("User-Centered Design and Pilot Testing of *myAURA* Through Focus Groups"). Alongside the focus groups, the participation of the EFA in all aims was instrumental in informing the user-centered design and development of the overall *myAURA* project according to stakeholder needs. The design and development included studying how social media can assist in predicting epilepsy outcomes,[28] human-centered dictionary refinement,[31] human-centered app design,[17,18] epilepsy-focused digital cohort extraction,[90] and our biomedical data science approach more broadly. Now that the data federation, KG construction, inference based on metric backbone sparsification, multilayer map visualization, human-centered design requirements, and pilot testing for *myAURA* have been completed—with constituent methods, tools, and open-source code—app production and deployment will continue in partnership with the EFA and other stakeholders.

To our knowledge, our team is the first to investigate PWEC practices and preferences, acquiring up-close and personal descriptions of the challenges they face, to seek out and curate epilepsy-related content. The resultant design framework is generalizable and useful for others interested in developing a similar app.[17,18] Indeed, the methods we detail here, and several of the data sources we have federated ("Data Federation and Processing"), are relevant not only to epilepsy patients but also to those with other chronic conditions.

Conclusion

Chronic health conditions unfold as a complex interplay among biological, psychological, and societal factors that change over time. Such complex multilayer dynamics of human health require new science, new tools, and new interdisciplinary thinking to accelerate data-driven discovery and management of chronic conditions.[102–104] We reported the advances our team has made in developing *myAURA*, a personal library application prototype and suite of methods to

support epilepsy research and self-management through the daunting array of treatments, drugs, interactions and side effects, diet, lifestyle, and stigma. We worked with PWEC and stakeholders to design and pilot-test the approach, which entailed federating many large-scale heterogeneous data streams into an epilepsy KG that we analyzed using novel network inference, sparsification, and visualization methods in support of personalized and explainable recommendation, digital cohort identification, and understanding of pharmacology in epilepsy. We showed that significant advances empowered by biomedical informatics are within reach for self-management and scientific discovery in epilepsy, especially by leveraging unconventional data from EHR, social media, and digital cohorts, along with computational and theoretical advances in characterizing and visualizing multilayer complex networks. We look forward to continuing to developing the *myAURA* system toward production and deployment of a full application for epilepsy and to expanding it to include a broad range of chronic conditions to benefit many more patients in the future.

Author contributions

Rion B. Correia (Data curation, Formal analysis, Investigation, Resources, Software, Supervision, Validation, Visualization), Jordan C. Rozum (Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Supervision, Validation, Visualization), Leonard Cross (Data curation, Software), Jack Felag (Data curation, Investigation, Software), Michael Gallant (Data curation, Software), Ziqi Guo (Data curation, Investigation, Software), Bruce W. Herr II (Data curation, Software, Visualization), Aehong Min (Data curation, Investigation, Validation), Jon Sanchez-Valle (Data curation, Resources, Software, Visualization), Deborah Stungis Rocha (Project administration, Supervision), Alfonso Valencia (Software, Supervision, Visualization), Xuan Wang (Data curation, Formal analysis, Investigation, Software, Validation), Katy Börner (Funding acquisition, Investigation, Methodology, Project administration, Supervision, Visualization), Wendy Miller (Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation), and Luis M. Rocha (Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Visualization)

Funding

This work was supported by National Institutes of Health, National Library of Medicine grant number 1R01LM012832. In addition, R.B.C. was partially funded by Fundação para a Ciência e a Tecnologia (grant PTDC/MEC-AND/30221/2017). L.M.R., K.B., and X.W. were partially funded by a National Science Foundation Research Traineeship "Interdisciplinary Training in Complex Networks and Systems" grant 1735095. L.M.R. was also partially funded by a Fulbright Commission fellowship and by Fundação para a Ciência e a Tecnologia (grant 2022.09122.PTDC, DOI: 10.54499/2022.09122.PTDC). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflicts of interest

The authors have no competing interests to declare.

Data availability

The data API, data aggregates, and code underlying this article are available through github at github.com/cns-iu/myaura for KG construction and github.com/CASCI-lab/for the backbone extraction pipeline, per technical, ethical, and legal considerations of human subject data and protected health information discussed in "Source Code, Software, and Data Sharing Policy".

References

1. EFA. Epilepsy statistics. 2016. Accessed June 15, 2024. http://www.epilepsy.com/learn/epilepsy-statistics.
2. Ngugi AK, Bottomley C, Kleinschmidt I, Sander JW, Newton CR. Estimation of the burden of active and life-time epilepsy: a meta-analytic approach. *Epilepsia*. 2010;51:883-890.
3. Hesdorffer DC, Beck V, Begley CE, et al. Research implications of the Institute of Medicine report, epilepsy across the spectrum: promoting health and understanding. *Epilepsia*. 2013;54:207-216.
4. Mölleken D, Richter-Appelt H, Stodieck S, Bengner T. Influence of personality on sexual quality of life in epilepsy. *Epileptic Disord*. 2010;12:125-132.
5. Tomson T, Surges R, Delamont R, Haywood S, Hesdorffer DC. Who to target in sudden unexpected death in epilepsy prevention and how? Risk factors, biomarkers, and intervention study designs. *Epilepsia*. 2016;57:4-16.
6. Miller WR, Young N, Friedman D, Buelow JM, Devinsky O. Discussing sudden unexpected death in epilepsy (SUDEP) with patients: practices of health-care providers. *Epilepsy Behav*. 2014;32:38-41.
7. Austin JK, Hesdorffer DC, Liverman CT, Schultz AM; Testimony Group. Testimonies submitted for the Institute of Medicine report: epilepsy across the spectrum: promoting health and understanding. *Epilepsy & Behavior*. 2012;25:634-661.
8. England MJ, Liverman CT, Schultz AM, Strawbridge LM. Epilepsy across the spectrum: promoting health and understanding: a summary of the Institute of Medicine report. *Epilepsy Behav*. 2012;25:266-276.
9. Bazargan-Hejazi S, Dehghan K, Edwards C, et al. The health burden of non-communicable neurological disorders in the USA between 1990 and 2017. *Brain Commun*. 2020;2:fcaa097.
10. Wood IB, Correia RB, Miller WR, Rocha LM. Small cohort of patients with epilepsy showed increased activity on Facebook before sudden unexpected death. *Epilepsy Behav*. 2022;128:108580.
11. Miller WR, Wion RK, Eads P. Evaluation of emergency department-based seizure and epilepsy education: exploring the need for early epilepsy self-management intervention. *Epilepsy Behav*. 2021;116:107702.
12. Majersik JJ, Ahmed A, Chen IHA, et al. A shortage of neurologists—we must act now: a report from the AAN 2019 Transforming Leaders Program. *Neurology*. 2021;96:1122-1134.
13. Elkhider H, Sharma R, Sheng S, et al. Predictors of no-show in neurology clinics. *Healthcare*. 2022;10:599.
14. Ross SC. An option for improving access to outpatient general neurology. *Neurol Clin Pract*. 2014;4:435-440.
15. Unger WR, Buelow JM. Hybrid concept analysis of self-management in adults newly diagnosed with epilepsy. *Epilepsy Behav*. 2009;14:89-95.
16. Miller WR. Patient-centered outcomes in older adults with epilepsy. *Seizure*. 2014;23:592-597.

17. Min A, Miller WR, Rocha LM, Börner K, Brattig Correia R, Shih PC. Just in time: challenges and opportunities of first aid care information sharing for supporting epileptic seizure response. *Proc ACM Hum-Comput Interact*. 2021;5:1-24.
18. Min A, Miller WR, Rocha LM, Börner K, Brattig Correia R, Shih PC. Understanding contexts and challenges of information management for epilepsy care. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems 2013*. Association for Computing Machinery; 2023:328. https://dx.doi.org/10.1145/3544548.3580949.
19. Simas T, Correia RB, Rocha LM. The distance backbone of complex networks. *J Complex Netw*. 2021;9:cnab021.
20. Steyvers M, Tenenbaum JB. The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cogn Sci*. 2005;29:41-78.
21. Johnson N, Rasmussen S, Joslyn C, Rocha L, Smith S, Kantor M. Symbiotic intelligence: self-organizing knowledge on distributed networks, driven by human interaction. In: *Proceedings of the 6th International Conference on Artificial Life*. MIT Press; 1998:403-407.
22. Börner K, Sanyal S, Vespignani A. Network science. *Ann Rev Inf Sci Technol*. 2007;41:537-607.
23. Wasserman S, Faust K. *Social Network Analysis: Methods and Applications*. Cambridge University Press; 1994.
24. Monge PR, Contractor NS. *Theories of Communication Networks*. Oxford University Press; 2003.
25. Barabási AL. *Linked: The New Science of Networks*. AAPT; 2003.
26. Chakrabarti D, Faloutsos C. Graph mining: laws, generators, and algorithms. *ACM Comput Surv*. 2006;38:2.
27. Börner K. Plug-and-play macroscopes. *Commun ACM*. 2011;54:60-69.
28. Correia RB, Wood IB, Bollen J, Rocha LM. Mining social media data for biomedical signals and health-related behavior. *Annu Rev Biomed Data Sci*. 2020;3:433-458.
29. Correia RB, Li L, Rocha LM. Monitoring potential drug interactions and reactions via network analysis of Instagram user timelines. *Pac Symp Biocomput*. 2016;21:492-503.
30. Correia RB. *Prediction of Drug Interaction and Adverse Reactions, with data from Electronic Health Records, Clinical Reporting, Scientific Literature, and Social Media, using Complexity Science Methods*. PhD thesis. Indiana University; 2019.
31. Min A, Wang X, Correia RB, Rozum J, Miller WR, Rocha LM. Refinement of an epilepsy dictionary through human annotation of health-related posts on Instagram. arXiv, arXiv:2405.08784, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.2405.08784
32. Wood IB, Varela PL, Bollen J, Rocha LM, Gonçalves-Sá J. Human sexual cycles are driven by culture and match collective moods. *Sci Rep*. 2017;7:17973.
33. Wood IB. *Time-Series Analysis of Sentiment in Social Media Can Predict Individual and Collective Behavior of Public Health Significance*. PhD thesis. Indiana University; 2023.
34. Davis CA, Ciampaglia GL, Aiello LM, et al. OSoMe: the IUNI observatory on social media. *PeerJ Comput Sci*. 2016;2:e87.
35. Sánchez-Valle J, Correia RB, Camacho-Artacho M, et al. Prevalence and differences in the co-administration of drugs known to interact: an analysis of three distinct and large populations. *BMC Med*. 2024;22:166.
36. Brattig Correia R, Araújo Kohler LP, Mattos MM, Rocha LM. City-wide electronic health records reveal gender and age biases in administration of known drug–drug interactions. *NPJ Digit Med*. 2019;2:74.
37. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res*. 2018;46:D1074-D1082.
38. Zarin DA, Tse T, Williams RJ, Califf RM, Ide NC. The ClinicalTrials.gov results database—update and key issues. *N Engl J Med*. 2011;364:852-860.
39. American Epilepsy Society. Find a doctor database. 2017. https://my.aesnet.org/FindaDoctor.
40. Sarker A, Ginn R, Nikfarjam A, et al. Utilizing social media data for pharmacovigilance: a review. *J Biomed Inform*. 2015;54:202-212.
41. Schwartz HA, Sap M, Kern ML, et al. Predicting individual well-being through the language of social media. *Pac Symp Biocomput*. 2016;21:516-527.
42. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. *Nucleic Acids Res*. 2016;44:D1075-D1079.
43. Lin FP, Anthony S, Polasek TM, Tsafnat G, Doogue MP. BICEPP: an example-based statistical text mining method for predicting the binary characteristics of drugs. *BMC Bioinformatics*. 2011;12:112.
44. US Food and Drug Administration. FDA Adverse Event Reporting System (FAERS). 2019. Accessed June 15, 2024. http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects.
45. US Food and Drug Administration. Medwatch: what is a serious adverse event. 2017. Accessed June 15, 2024. https://www.fda.gov/safety/reporting-serious-problems-fda/what-serious-adverse-event.
46. MedDRA. Medical dictionary for regulatory activities. Accessed November 19, 2021. meddra.org.
47. Kolchinsky A, Lourenço A, Li L, Rocha LM. Evaluation of linear classifiers on articles containing pharmacokinetic evidence of drug-drug interactions. *Pac Symp Biocomput*. 2013;18:409-420.
48. Kolchinsky A, Lourenço A, Wu HY, Li L, Rocha LM. Extraction of pharmacokinetic evidence of drug-drug interactions from the literature. *PLoS One*. 2015;10:e0122199.
49. Wu HY, Shendre A, Zhang S, et al. Translational knowledge discovery between drug interactions and pharmacogenetics. *Clin Pharmacol Ther*. 2020;107:886-902.
50. Zhang S, Wu H, Wang L, et al. Translational drug–interaction corpus. *Database*. 2022;2022:baac031.
51. Bradley M, Lang P. *Affective Norms for English Words (ANEW): Technical Manual and Affective Ratings*. Technical Report C-1. University of Florida; 1999.
52. Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. In: *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, Ann Arbor, Michigan, USA*; 2014.
53. Tausczik YR, Pennebaker JW. The psychological meaning of words: LIWC and computerized text analysis methods. *J Lang Soc Psychol*. 2010;29:24-54.
54. Simas T, Rocha LM. Distance closures on complex networks. *Net Sci*. 2015;3:227-268.
55. Abi-Haidar A, Kaur J, Maguitman A, et al. Uncovering protein interaction in abstracts and text using a novel linear model and word proximity networks. *Genome Biol*. 2008;9:S11.
56. Jaccard P. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles*. 1901;37:241-272.
57. Grefenstette G. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic; 1994.
58. Ciampaglia GL, Shiralkar P, Rocha LM, Bollen J, Menczer F, Flammini A. Computational fact checking from knowledge networks. *PLoS One*. 2015;10:e0128193.
59. Baeza-Yates R, Ribeiro-Neto B. *Modern Information Retrieval*. ACM Press, Addison-Wesley; 1999.
60. Turney PD. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In: *Proceedings of the European Conference on Machine Learning*. Springer Berlin Heidelberg; 2001:491–502.
61. Klir GJ. *Uncertainty and Information: Foundations of Generalized Information Theory*. Wiley-IEEE Press; 2005.
62. Kalavri V, Simas T, Logothetis D. The shortest path is not always a straight line: leveraging semi-metricity in graph analysis. *Proc VLDB Endow*. 2016;9:672-683.

63. Wang S, Zhou W, Jiang C. A survey of word embeddings based on deep learning. *Computing*. 2020;102:717-740.
64. Selva Birunda S, Devi KR. A review on word embedding techniques for text classification. In: *Innovative Data Communication Technologies and Application: Proceedings of ICIDCA*. Vol. 2021. Springer; 2020:267-281. https://dx.doi.org/10.1007/978-981-15-9651-3
65. Han H, Liu X. The challenges of explainable AI in biomedical data science. *BMC Bioinformatics*. 2022;22:443.
66. Rocha LM. Semi-metric behavior in document networks and its application to recommendation systems. In: Loia V, ed. *Soft Computing Agents: A New Perspective for Dynamic Information Systems*. IOS Press; 2002:137-163.
67. Rocha LM, Simas T, Rechtsteiner A, Giacomo MD, Luce R. MyLibrary@LANL: proximity and semi-metric networks for a collaborative and recommender web service. In: *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*. IEEE Press; 2005:565-571.
68. Simas T, Rocha LM. Semi-metric networks for recommender systems. In: *Proceedings of the 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology WI-IAT'12*. IEEE Computer Society; 2012:175-179.
69. Verspoor K, Cohn J, Joslyn C, et al. Protein annotation as term categorization in the gene ontology using word proximity networks. *BMC Bioinformatics*. 2005;6:S20.
70. Kolchinsky A, Abi-Haidar A, Kaur J, Hamed AA, Rocha LM. Classification of protein-protein interaction full-text documents using text and citation network features. *IEEE/ACM Trans Comput Biol Bioinform*. 2010;7:400-411.
71. Lourenço A, Conover M, Wong A, et al. A linear classifier based on entity recognition tools and a statistical approach to method extraction in the protein-protein interaction literature. *BMC Bioinformatics*. 2011;12:S12.
72. Manz T, Gold I, Patterson NH, et al. Viv: multiscale visualization of high-resolution multiplexed bioimaging data on the web. *Nat Methods*. 2022;19:515-516.
73. Börner K. *Atlas of Forecasts: Modeling and Mapping Desirable Futures*. MIT Press; 2021.
74. Azoulay P, Graff-Zivin J, Uzzi B, et al. Toward a more scientific science. *Science*. 2018;361:1194-1197.
75. Börner K. *Atlas of Knowledge: Anyone Can Map*. MIT Press; 2015.
76. Ginda M, Herr BW, Börner K, et al. Introducing the open biomedical map of science. *Front Res Metr Anal*. 2023;8:1274793.
77. Correia RB, Almeida JM, Wyrwoll MJ, et al. The conserved genetic program of male germ cells uncovers ancient regulators of human spermatogenesis. *Elife*. 2024;13:RP95774.
78. Jacobsen A, de Miranda Azevedo R, Juty N, et al. FAIR principles: interpretations and implementation considerations. *Data Intell*. 2020;2:10-29. https://dx.doi.org/10.1162/dint_r_00024
79. Sanchez-Valle J. DDInteract. 2024. Accessed June 15, 2024. http://disease-perception.bsc.es/ddinteract/.
80. Wang X, Correia RB, Wood IB, et al. Systematic prediction of drug-drug-interaction study types and discovery of evidence gaps in the literature. 2024; submitted.
81. Costa FX, Correia RB, Rocha LM. The distance backbone of directed networks. In: Complex Networks and their Applications XI: Proceedings of the Eleventh International Conference on Complex Networks and their Applications: Complex Networks 2022. Vol. 2. Springer; 2023:135-147. https://dx.doi.org/10.1007/978-3-031-21131-7
82. Galvin F, Shore SD. Distance functions and topologies. *Am Math Monthly*. 1991;98:620-623.
83. Dijkstra E. A note on two problems in connexion with graphs. *Numer Math*. 1959;1:269-271.
84. Brattig Correia R, Barrat A, Rocha LM. Contact networks have small metric backbones that maintain community structure and are primary transmission subgraphs. *PLoS Comput Biol*. 2023;19:e1010854.
85. Soriano Paños D, Costa FX, Rocha LM. Semi-metric topology characterizes epidemic spreading on complex networks. arXiv, arXiv:2311.14817, preprint: not peer reviewed.
86. Dorsant-Ardon V, Sanjay AB, Rocha LM, Correia RB, Apostolova LG. Gene co-expression network analyses in mild cognitive impairment. *Alzheimers Dement*. 2023;19:e082238.
87. Simas T, Suckling J. Commentary: semi-metric topology of the human connectome: sensitivity and specificity to autism and major depressive disorder. *Front Neurosci*. 2016;10:353.
88. Peeters S, Simas T, Suckling J, et al.; for Genetic Risk and Outcome of Psychosis (GROUP). Semi-metric analysis of the functional brain network: relationship with familial risk for psychotic disorder. *NeuroImage: Clin*. 2015;9:607-616.
89. Team CASCI. DistanceClosure. 2024. Accessed June 15, 2024. https://github.com/CASCI-lab/distanceclosure.
90. Guo Z, Felag J, Rozum JC, Correia RB, Rocha LM. Selecting focused digital cohorts from social media using the metric backbone of biomedical knowledge graphs. arXiv, arXiv:2405.07072, preprint: not peer reviewed.
91. Perry BL, Pescosolido BA, Small ML, McCranie A. Introduction to the special issue on ego networks. *Net Sci*. 2020;8:137-141.
92. CNS Team. MyAura-specific knowledge graph visualization tool. 2023. Accessed June 15, 2024. https://cns-iu.github.io/myaura/.
93. Borner k. map4sci visualization suite. 2023. Accessed June 15, 2024. https://github.com/cns-iu/map4sci/.
94. DeLuca F, Hossain I, Kobourov S, Börner K. Multi-level tree based approach for interactive graph visualization with semantic zoom. arXiv, arXiv:1906.05996. 2019. preprint: not peer reviewed.
95. Ahmed R, Angelini P, Bekos MA, et al. Splitting vertices in 2-layer graph drawings. *IEEE Comput Graph Appl*. 2023;43:24-35.
96. Saket B, Scheidegger C, Kobourov S, Börner K. Map-based visualizations increase recall accuracy of data. *Comput Graph Forum*. 2015;34:441-450.
97. Laugwitz B, Held T, Schrepp M. Construction and evaluation of a user experience questionnaire. In: Holzinger A, ed. *HCI and Usability for Education and Work*. Springer Berlin Heidelberg; 2008:63-76.
98. Beddiar DR, Oussalah M, Seppänen T. Automatic captioning for medical imaging (MIC): a rapid review of literature. *Artif Intell Rev*. 2023;56:4019-4076.
99. Edge D, Trinh H, Cheng N, et al. From local to global: a graph rag approach to query-focused summarization. arXiv, arXiv:2404.16130, preprint: not peer reviewed.
100. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv Neural Inf Process Syst*. 2020;33:9459-9474.
101. Dihan QA, Nihalani BR, Tooley AA, Elhusseiny AM. Eyes on Google's NotebookLM: using generative AI to create ophthalmology podcasts with a single click. *Eye*. 2024:1-2.
102. Pescosolido BA, Olafsdottir S, Sporns O, et al. The social symbiome framework: linking genes-to-global cultures in public health using network science. In: Neal Z, ed. *Handbook of Applied System Science*. Taylor and Francis Inc.; 2016:25-48.
103. Trochim WM, Cabrera DA, Milstein B, Gallagher RS, Leischow SJ. Practical challenges of systems thinking and modeling in public health. *Am J Public Health*. 2006;96:538-546.
104. Rusoja E, Haynie D, Sievers J, et al. Thinking about complexity in health: a systematic review of the key systems thinking and complexity ideas in health. *J Eval Clin Pract*. 2018;24:600-606.

Journal of the American Medical Informatics Association, 2026, 33, 167–181
https://doi.org/10.1093/jamia/ocaf012
Research and Applications