Journal of the American Medical Informatics Association, 24(1), 2017, 13–23
doi: 10.1093/jamia/ocw003
Advance Access Publication Date: 17 May 2016
Research and Applications

Research and Applications

Learning disease relationships from clinical drug trials

Bryan Haslam[1] and Luis Perez-Breva[2]

[1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA and [2]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Correspondence to Bryan Haslam, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 77 Massachusetts Ave E70-1263, Cambridge, MA 02139, USA; bhaslam@mit.edu.

Received 11 August 2015; Revised 23 December 2015; Accepted 3 January 2016

ABSTRACT

Objective: Our objective is to test the limits of the assumption that better learning from data in medicine requires more granular data. We hypothesize that clinical trial metadata contains latent scientific, clinical, and regulatory expert knowledge that can be accessed to draw conclusions about the underlying biology of diseases. We seek to demonstrate that this latent information can be uncovered from the whole body of clinical trials.

Materials and Methods: We extract free-text metadata from 93 654 clinical drug trials and introduce a representation that allows us to compare different trials. We then construct a network of diseases using only the trial metadata. We view each trial as the summation of expert knowledge of biological mechanisms and medical evidence linking a disease to a drug believed to modulate the pathways of that disease. Our network representation allows us to visualize disease relationships based on this underlying information.

Results: Our disease network shows surprising agreement with another disease network based on genetic data and on the Medical Subject Headings (MeSH) taxonomy, yet also contains unique disease similarities.

Discussion and Conclusion: The agreement of our results with other sources indicates that our premise regarding latent expert knowledge holds. The disease relationships unique to our network may be used to generate hypotheses for future biological and clinical research as well as drug repurposing and design. Our results provide an example of using experimental data on humans to generate biologically useful information and point to a set of new and promising strategies to link clinical outcomes data back to biological research.

Key words: machine learning, networks, clinical trials

BACKGROUND AND SIGNIFICANCE

The collection of all clinical trials represents research carried out on millions of humans[1] along with thousands of decisions made by researchers, clinicians, and executives managing billions of dollars.[2] Each trial teaches us something about the specific relationship between an indication and an intervention that can be summarized in terms of safety and efficacy. However, the question of how to learn from the collection of clinical trials remains open. Little is known about the information clinical trials reveal as a whole.

We hypothesize that, at present, different trials are only connected indirectly through the expertise accumulated by the scientific, clinical, regulatory, and executive decision-makers who deployed their expertise in those trials. We believe this expertise is latent in the collection of clinical trials and can be accessed via inference. Viewing the collection of clinical trials as observables emerging from this latent knowledge, we show how to leverage clinical trial metadata to arrive at unique insights into the relationships between diseases. We further show how these insights may be useful in linking clinical data back to biology by generating hypotheses for future biological research.

Collections of clinical trials are now available in clinical trials registries, enabling work like ours to benefit from multiple clinical

trials together.[3] The desire to increase learning from multiple trials and possibly further inform scientific research has led many researchers to push for greater access to patient-level clinical trial data and even electronic health records.[4–6] The prevailing assumption within the field is that the more granular the data, the better the learning.

In this paper we explore whether relationships among diseases can be learned from superficial data, such as the kind of metadata available in clinical trial registries, and how the relationships uncovered compare to relationships learned from molecular and biological data.

The notion that new knowledge may be generated from the structured aggregation of databases has been used before to explore relationships between diseases or between drugs. For example, Goh et al.[7] and Butte et al.[8] compared diseases and phenotypes by shared gene correlations based on the Online Mendelian Inheritance in Man database and the Gene Expression Omnibus database. Zhou et al.[9] compared diseases by shared symptoms from text in PubMed. Hidalgo et al.[10] compared diseases by shared comorbid conditions. Yildirim et al.[11] compared drugs based on shared targets. Campillos et al.[12] and Tatonetti et al.[13] compared drugs based on shared side effects. These examples provide a strategy to access disease similarities that would not have been apparent from individual experiments.[14]

New insights into disease similarities or dissimilarities can have dramatic results. For example, the similarity between psoriasis and multiple sclerosis led to the blockbuster drug Tecfidera (dimethyl fumarate) being used to treat relapsing-remitting multiple sclerosis.[15] As another example, the discovery of genetic dissimilarity in breast cancer corresponding to dissimilar prognoses led to new and improved diagnosis and treatment options.[16] Similarity metrics between drugs have been used to predict indications, targets, and drug interactions.[17–19] Such predictions have enabled repurposing of existing drugs and avoidance of adverse drug reactions.[20]

To some degree, all of these examples may be viewed as accessing latent information about underlying biology. In this work we follow an approach similar in spirit, but introduce the hypothesis that the underlying biology is contained in the expert knowledge used in deciding to conduct the trials. Expert knowledge may come from literature reviews, proprietary biological experiments, other clinical trials, or the summaries of expert opinions. We further hypothesize that such expert knowledge may be accessed as latent information from clinical trial metadata. By uncovering that latent information, it is possible to extract what experts collectively know about the relationships among tested diseases. Using this information, one may explore the relationships among diseases and generate hypotheses to guide future research. In this work we use aggregated clinical trial metadata to construct relationships among diseases and show that similar relationships can be found when compared to analyses based on detailed biological data.

To the best of our knowledge, studying disease relationships by using the entire set of clinical drug trials along with the premise of expert knowledge latent in the trial metadata is novel, and it is a key contribution of this work. Our approach and subsequent results imply that there is much more knowledge to be gained from existing clinical trial data. It also suggests a path to compare patient-level data across different trials when such data becomes more widely available.

The paper is organized as follows: We explain how we used free-text metadata from drug trials on ClinicalTrails.gov to construct a model of the diseasome. We then explore the connectivity between diseases and drugs and visualize the data in a network layout. We report on the validation of the disease-disease network (DDN) against a standard disease taxonomy and a diseasome built from genetic data. The relationships derived from our network show surprising agreement with relationships based on genetic data or medical taxonomies and show promise for informing future scientific research.

METHODS

Construction of the disease-disease network

We extracted metadata from 93 654 clinical drug trials on ClinicalTrials.gov (see Supplementary Methods). The metadata included a list of 1 or more free-text strings for conditions (diseases) and a similar list for drugs. Comparing diseases or drugs from different trials was sometimes ambiguous, because 2 different text strings could represent the same concept. For example, there are 73 strings that represent the single concept type 2 diabetes mellitus (Supplementary Table 1).

Some resources have already been constructed to disambiguate diseases and drugs, such as the option to browse trials by condition or drug intervention on ClinicalTrials.gov[21] and use of the Aggregate Analysis of ClincalTrials.gov (AACT) database.[22] These resources included errors, such as imidacloprid listed as a drug in 129 trials when browsing by intervention on ClinicalTrials.gov and in 6 trials in the AACT database. Imidacloprid is actually an insecticide that is not tested in any trials listed on ClinicalTrials.gov. We traced this false positive to the trade name Advantage, which is used as a synonym for imidacloprid in the 2013 edition of Medical Subject Headings (MeSH). These false positives occur because both resources rely on an automatic algorithm for finding MeSH terms.[22] Such false positives would erroneously connect diseases in our analysis. We also found that these resources have built-in inferences based on a National Library of Medicine (NLM) algorithm or annotations by clinicians. Our goal was to use the raw data as much as possible, without introducing layers of inference.

Much of the work that analyzes large sets of clinical trials is based on the AACT database.[23–25] Other work focuses on specific aspects of the trials, such as drug combinations[26] or participation criteria,[27,28] which use algorithms tailored to the specific fields to automatically extract their datasets. Due to the nature of our approach, it was more important for us to reduce false positives, and we therefore curated the data manually. To enable comparison, we built a thesaurus of terms, starting with diseases (Figure 1). We started with the MeSH vocabulary[29] as a base thesaurus. Another option was UMLS,[30] which is a more comprehensive thesaurus based on many databases and can be accessed with enhanced tools such as MetaMap.[31] We chose MeSH because contributors to ClinicalTrials.gov are encouraged to use its vocabulary,[32] and by using a single database we avoided inferences as described above. MeSH only identified 22% of the unique disease strings listed in ClinicalTrials.gov, accounting for 56% of all disease strings and 70% of the trials. We manually reviewed the 17 970 disease strings not identified by MeSH, comparing each one to the 20 closest terms in MeSH generated by fuzzy string matching (see Supplementary Methods for details). If a match could not be made and the disease string occurred repeatedly, we created a new "data-derived" term in the thesaurus. The number of MeSH disease terms used over time peaked in 2008, and the percentage of terms we added to our thesaurus increased linearly over time (Supplementary Figure 1), suggesting that the 2014 MeSH vocabulary did not include terms

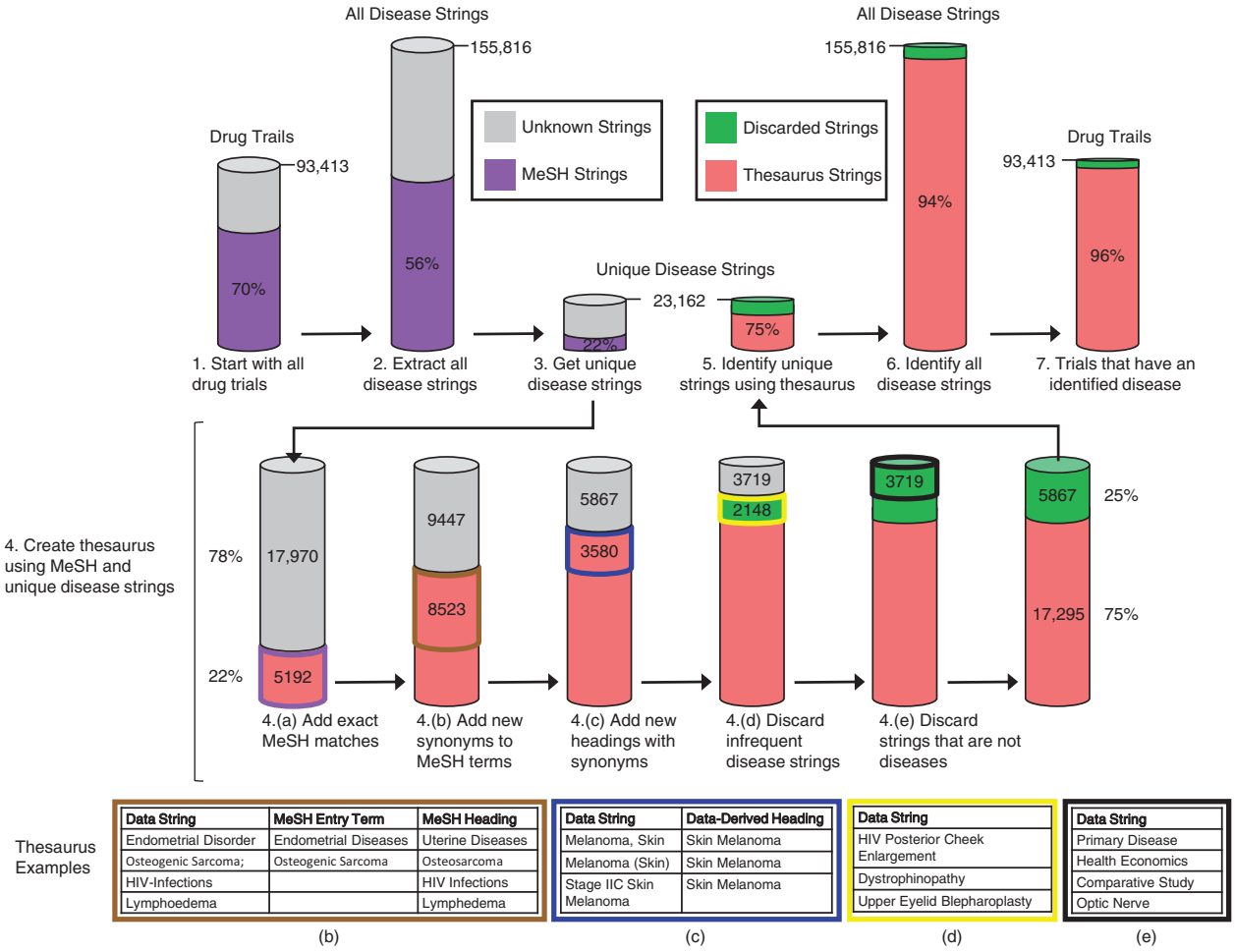

Figure 1. Creating a thesaurus maximizes the data that can be used. To compare disease terms to each other, we needed a standard vocabulary with synonyms. We started with the Medical Subject Headings (MeSH), but only 70% of drug trials on ClinicalTrials.gov and 56% of the diseases listed in those trials could be found in MeSH. We augmented MeSH by looking at every unique disease string, of which only 22% are in MeSH. Going through the remaining 78% manually, we either added another synonym to a MeSH term (4b), created new terms from the data with accompanying synonyms (4c), or discarded infrequent or irrelevant strings (4d and 4e). Every unique string was reviewed and either included in our thesaurus or discarded. From our thesaurus we identified 94% of all disease strings, enabling us to compare data from 96% of the trials.

currently used in research, hence the need for our enhanced thesaurus. Using our thesaurus, we identified 94% of the disease strings for 96% of the drug trials listed in ClinicalTrials.gov.

Drugs were more challenging to disambiguate, because one drug string often included multiple drugs and because experimental drugs were often not found in MeSH. Of 63 066 unique drug strings, we generated 503 270 possible substrings and used automated and manual filtering to identify all drugs and construct a thesaurus (see Supplementary Methods and Supplementary Figure 2). We analyzed the drug thesaurus and found patterns similar to that of the disease thesaurus (Supplementary Figures 3 and 4). We then used the drug thesaurus to identify individual drugs in drug strings (Supplementary Table 2). To assess accuracy, we took a sampling of 100 random trials with 216 drugs and found that 98% were identified (Supplementary Table 3). We also sampled 100 drug strings that were not mapped to any drugs and found that 92% were correctly excluded (Supplementary Table 4).

The resulting disease-drug dataset accounts for 93 069 trials and includes 132 822 diseases (3663 unique) and 175 584 drugs (7349 unique). HIV infection was the most tested disease, with 2772 trials involving 700 drugs. There were 1211 unique diseases tested in more than 10 trials with at least 1 drug and 1784 unique drugs tested in more than 10 trials with at least 1 disease. The variety of trials involving different diseases and drugs provides the connectivity for a network.

To construct the network, we started with the disease-drug network, a bipartite network of disease nodes and drug nodes (Figure 2). Diseases and drugs are linked by trials, with the width of the edge proportional to the number of trials with both the disease and the drug. The disease-disease network (DDN) is constructed using only diseases as nodes and the edges representing the number of drugs tested at least once on both diseases. The connectivity of the disease-drug network follows a power law (Supplementary Figure 5A), while the connectivity of the DDN follows an exponential distribution (Supplementary Figure 5B).[33] We assigned the weight of all edges in the graph to 1, indicating a binary connection between diseases, for simplicity in this work.

Visualization of the disease-disease network

Visualizing the network is difficult, because there are 3663 disease nodes with hundreds of thousands of edges between them. To reduce

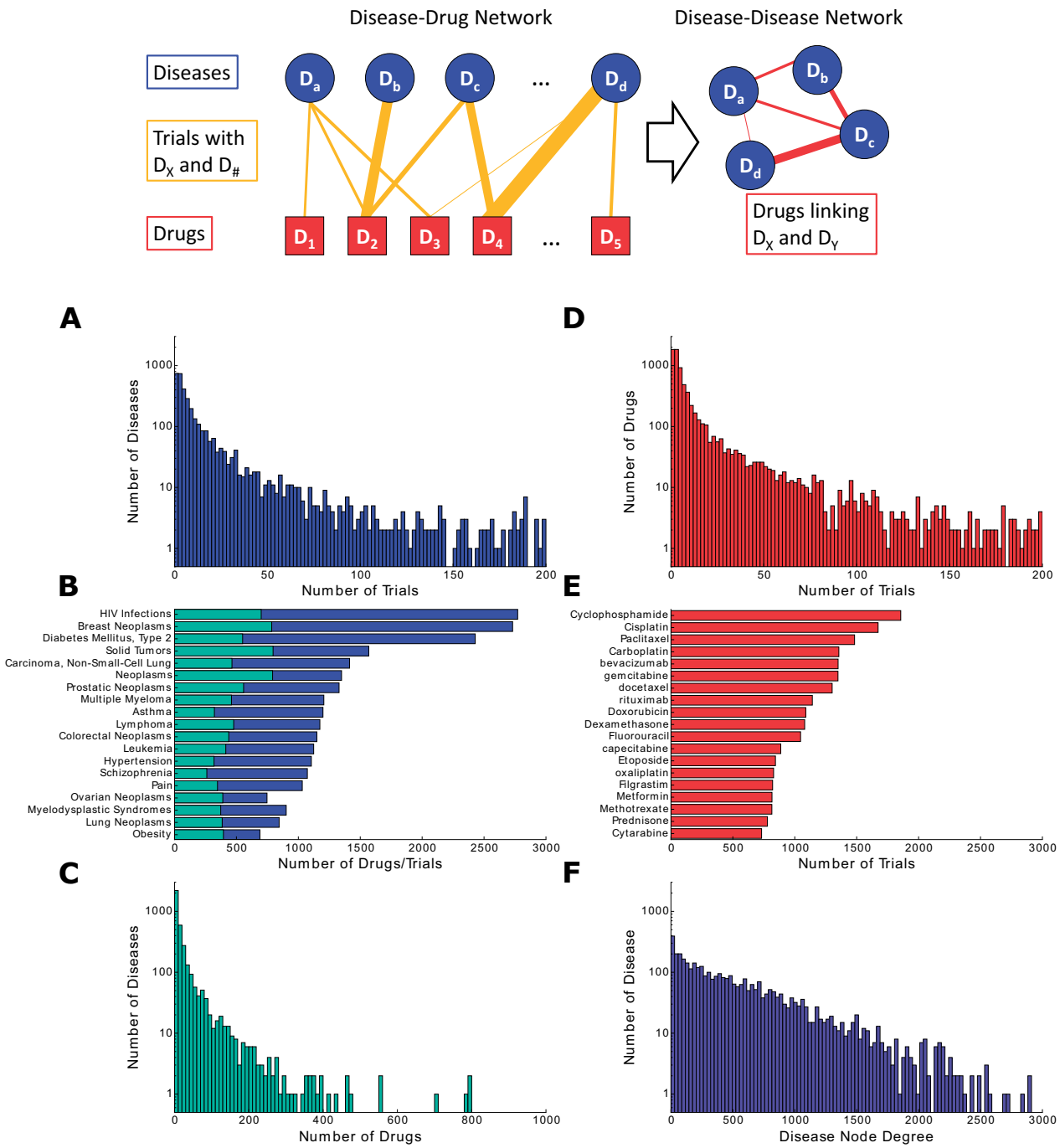

Figure 2. Construction of the disease-disease network (DDN) with descriptive statistics. The DDN was constructed from a bipartite network of diseases and drugs linked by trials. In the bipartite network, the thickness of edges corresponds to the number of trials that have both the disease and the drug the edge connects. In the DDN, diseases are linked by drugs, with the edges proportional to the number of drugs tested in trials with both diseases the edge connects. (**A**) The number of diseases that occured in a given number of trials. Diseases appear in thousands of trials, but for visualization purposes the plot was truncated at 200 trials. Note that the Y-axis in each plot is a logarithmic scale. (**B**) The top 15 diseases by number of trials (darker/longer bars) and number of drugs (lighter/shorter bars). (**C**) The number of diseases tested with a given number of unique drugs. (**D**) The number of drugs that occur in a given number of trials. (**E**) The top 15 drugs by number of trials. (**F**) The number of diseases associated with a given number of other diseases by the criteria that both diseases were tested with a particular drug. The *X*-axis can be viewed as the degree of each node in a network of diseases linked to each other.

the number of edges, we filtered edges to keep ones with strong relationships. We defined an edge as strong if 1 of the 2 disease nodes it connects is frequently associated with the drug the edge represents. A frequently associated disease for a given drug is one that shows up in a significant percentage of all trials for that drug. To determine significance, we used a binary test with a cutoff *P*-value of 0.001 and the Bonferonni correction for comparing multiple diseases. We selected this method for filtering because of a pattern we found in drug trials. New diseases tested on a drug are either tested in conjunction with an established indication for that drug or tested on completely new diseases. We hypothesized that the first case suggests a deeper characterization of the drug and diseases, while the latter case suggests an

exploration of possible new indications. We treated the diseases in the first case as having a strong relationship. Our filtering for strong relationships between diseases is only one way to examine the data. In the discussion, we explain why subtle weak relationships are potentially more interesting.

The network graph was laid out using the Fruchterman and Reingold method, though other force-directed layout algorithms gave similar results (Supplementary Methods and Supplementary Figure 6). Node size is proportional to the number of drugs tested on the disease, and edge width is proportional to the number of drugs tested on both diseases. Nodes are colored according to MeSH categories of diseases (Supplementary Methods and Supplementary Table 5).

Validation with the MeSH taxonomy

The MeSH disease taxonomy was constructed by experts based on biological and clinical understanding. If the DDN can reproduce the MeSH taxonomy, it would suggest that the DDN captures the same level of information implicitly that experts explicitly outlined when constructing MeSH. To explore similarities between the DDN and MeSH, we quantitatively evaluated clustering of diseases in the DDN by MeSH category. First, we evaluated the internal consistency of clusters in our network visualization using the nearest neighbor index.[34] Second, we evaluated how distinct clusters are based on graph theoretic distance. Third, we evaluated how consistent the DDN and MeSH are compared to a randomly constructed network using a binomial test.

Validation with the human disease network

We also validated the DDN by comparing it to the human disease network (HDN), which was constructed using a database of disease-gene associations.[7] The HDN was validated by examining clustering by disease categories that match the MeSH categories we used (Supplementary Table 5). First, we evaluated the internal consistency of clusters of nodes within the same category by measuring the fraction of edges connecting nodes within that category. Second, we used the ratio of shortest paths within versus without of a category to derive a graph theoretic measure of clustering for each disease category. Third, we evaluated how much overlap there was relative to a randomly constructed network using the binomial test.

Prediction potential

We tested the potential of the DDN for prediction by building a rudimentary recommender engine for clinical trials. We used the entire unfiltered DDN, rather than the filtered version that we used for visualization purposes, to capture the subtle relationships among diseases and not just the strongest ones. Our training set contained trials starting before 2011 and our test set contained trials starting in 2011 or later. The dataset contains 2160 diseases that were tested with at least 1 drug in the training set and 1 drug in the test set. There are 7349 possible drugs to predict, with 54 509 disease-drug pairs in the training set and 19 157 disease-drug pairs in the test set. Each disease is represented by a vector of drug variables, with 1 indicating that the drug was tested in a trial with the disease and 0 otherwise. The purpose of the recommender engine is to suggest drugs that had not previously been tested on a given disease but may be relevant to a disease based on data in the training set. For a given disease, we made predictions about each drug using collaborative filtering,[35] with a cosine similarity metric or the normalized inner product between disease vectors. We evaluated the performance of the recommender engine looking at the area under the ROC curve[36] for each disease in the 3.5-year period after 2011.

RESULTS

The disease-disease network

The network graph resulting from our layout (Figure 3) contains a giant component, with 1101 nodes and 6972 edges. The distance between nodes in the graph represents similarity based on shared drugs directly or through other disease nodes. By visual inspection, clustering of nodes of the same color or MeSH category can be seen. Many nodes in close proximity were expected, such as Crohn's Disease and Ulcerative Colitis or Parkinson's and Alzheimer's disease. At the same time there were surprises, such as hypertension and Parkinson's being close together. We also observed similarities of disease categories, such as psychiatric and nervous system diseases next to each other or cardiovascular and metabolic diseases mixed together.

Validation with MeSH

Validation of the DDN against the Medical Subject Headings (MeSH) taxonomy is shown in Figure 4. The nearest neighbor index (NNI) for a group of data points in a plane indicates whether the points are randomly spaced (an index of 1), nonrandomly clustered (an index smaller than 1), or non-randomly spaced apart (an index larger than 1). The NNI for each MeSH category (Figure 4A) is less than 1 for all categories except Skin. Compared to randomly connected networks, 13 of the 15 disease categories with 10 or more nodes had a significantly smaller ($P < .05$) NNI (Supplementary Table 7). All 3 disease categories with fewer than 10 nodes were not significant. All P-values were calculated empirically using Monte Carlo simulations (Supplementary Table 7).[37]

Figure 4B shows how close nodes of the same MeSH category (colored bars) are compared to how close nodes of different MeSH categories (gray bars) are, using shortest path distance. Distinct clusters have a significantly shorter colored bar than gray bar, which is the case for all 15 disease categories with 10 or more nodes. Compared to randomly connected networks, 14 of these categories have a significantly ($P < .05$) shorter colored bar (Supplementary Table 7). Only 1 of the 3 disease categories with fewer than 10 nodes had a shorter colored bar, which was also significant compared to randomly connected networks. Figure 4C shows the binomial distribution of the number of edges connecting nodes of the same category if they were randomly placed in the network, and the red arrow indicates how many correct links we observed in the DDN. For the binomial test, the P-value is too small to be be calculated using double floating point precision (Supplementary Methods).

The 3 evaluations show that the connectivity of the network as a whole significantly reflects categories in the MeSH. In addition to the clustering within categories, we also noted related diseases of different categories that are close, such as AIDS (Endocrine) and Hepatitis C (Digestive), Myelodysplastic Syndromes (Hemic and Lymphatic) and Leukemia (Neoplasms), and Hypercholesterolemia (Metabolic) and several cardiovascular diseases.

Validation with the HDN

There are visual similarities between the DDN and the Human-Disease Network (HDN), such as neoplasms/cancer being the largest cluster. There are differences, too, such as deafness being prominent

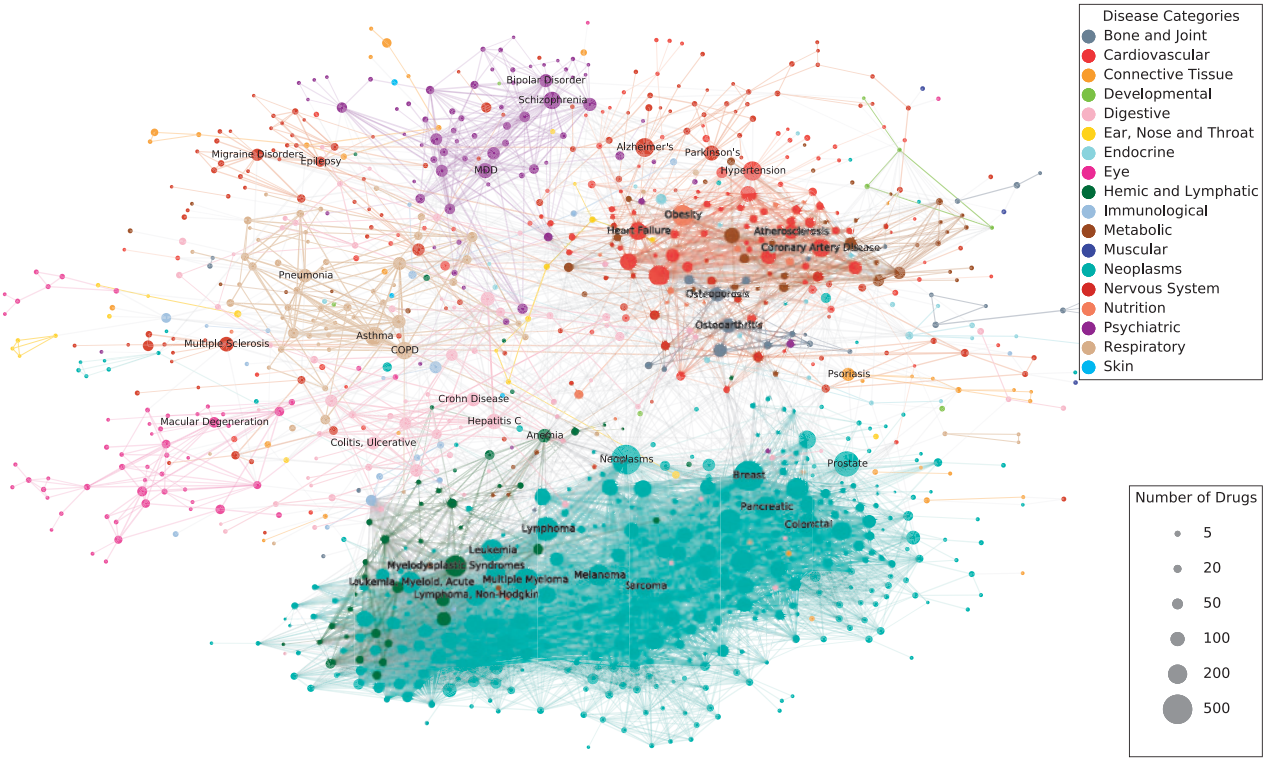

Figure 3. Visualization of the disease-disease network (DDN). In the DDN, diseases are represented as nodes, with the size of the node proportional to the number of drugs tested on that disease. Edges between nodes represent drugs tested on both diseases the edge is connected to. Thickness of the edge is proportional to the number of drugs tested on the 2 diseases. For ease of comparison with MeSH, we colored nodes according to MeSH disease subtrees, though that information was not used by the visualization algorithm. A cluster of nodes of 1 color indicates the DDN captures information about the relationships between diseases that can be found in the MeSH taxonomy.

in the HDN but absent in our plot, which reflects the different data sources. Deafness may be strongly associated with certain genes, but it does not currently have pharmaceutical treatment options.

Quantitative validation of the DDN against the HDN is shown in Figure 5. The average degree fraction within a disease category indicates how connected the nodes in the category are to each other. Figure 5A shows that average degree fraction is similar or larger for the DDN (colored bars) compared to the HDN (gray bars) for 13 of the 15 categories with 10 or more nodes and for 0 of the 3 categories with fewer than 10 nodes. The 2 categories with more than 10 nodes and a smaller average degree fraction in the DDN are Metabolic and Hemic and Lymphatic. Hemic and Lymphatic is an interesting case, where the DDN has a smaller degree fraction than the HDN. This is true because it is mixed with Neoplasms in the DDN, which may reflect the similarity in treatment in hematology and oncology, while the genetic basis may be more distinct. The difference between the DDN and HDN compared to the difference between randomly connected networks and the HDN is significant ($P < .01$) for all categories except Muscular and Skin, which do not have any directly connected nodes (Supplementary Table 8).

Taking into account indirect connections between nodes, Figure 5B shows the mean of the ratio of the shortest path within versus without for each category, where a smaller ratio indicates tighter clustering within the category. The DDN has a similar or smaller ratio for 10 of the 15 categories with 10 or more nodes and for 1 of the 3 categories with fewer than 10 nodes. The difference between the DDN and HDN compared to the difference between randomly connected networks and the HDN is significant ($P < .01$) for all categories except Connective Tissue, Muscular, and Skin (Supplementary Table 8).

Comparing the 2 networks directly, we found 181 common nodes with 764 edges among those nodes in the DDN and 192 edges among the same nodes in the HDN. The expected number of overlapping edges is a binomial distribution (Figure 5C). We observed that 73 edges were the same, giving a P-value of 9×10^{-42} (see Supplementary Methods). There is significant overlap in disease relationships found in the DDN compared to the DDN, even though the 2 networks were constructed using very different datasets.

Prediction potential

We plotted the AUC for all diseases as a scatter plot also showing the disease category, the number of drugs in the training set, and the number of drugs in the test set (Figure 6). Random predictions would give an AUC around 0.5, while the majority of our predictions had an AUC much larger. The average AUC for diseases was 0.845. The histogram of all AUCs compared to random predictions benchmarks the predictive ability of the network (Supplementary Figure 7). Using the Shapiro-Wilks test for normality of the AUC scores, the P-value is 6×10^{-39}. We note that if a trial did not occur in the 3.5-year test set time period, this did not indicate that a trial will not happen in the future or that there is no connection between the drug and the disease, so this result represents a conservative estimate. Examples of diseases with an AUC of more than 0.95 are provided in Supplementary Table 9, along with references to recent

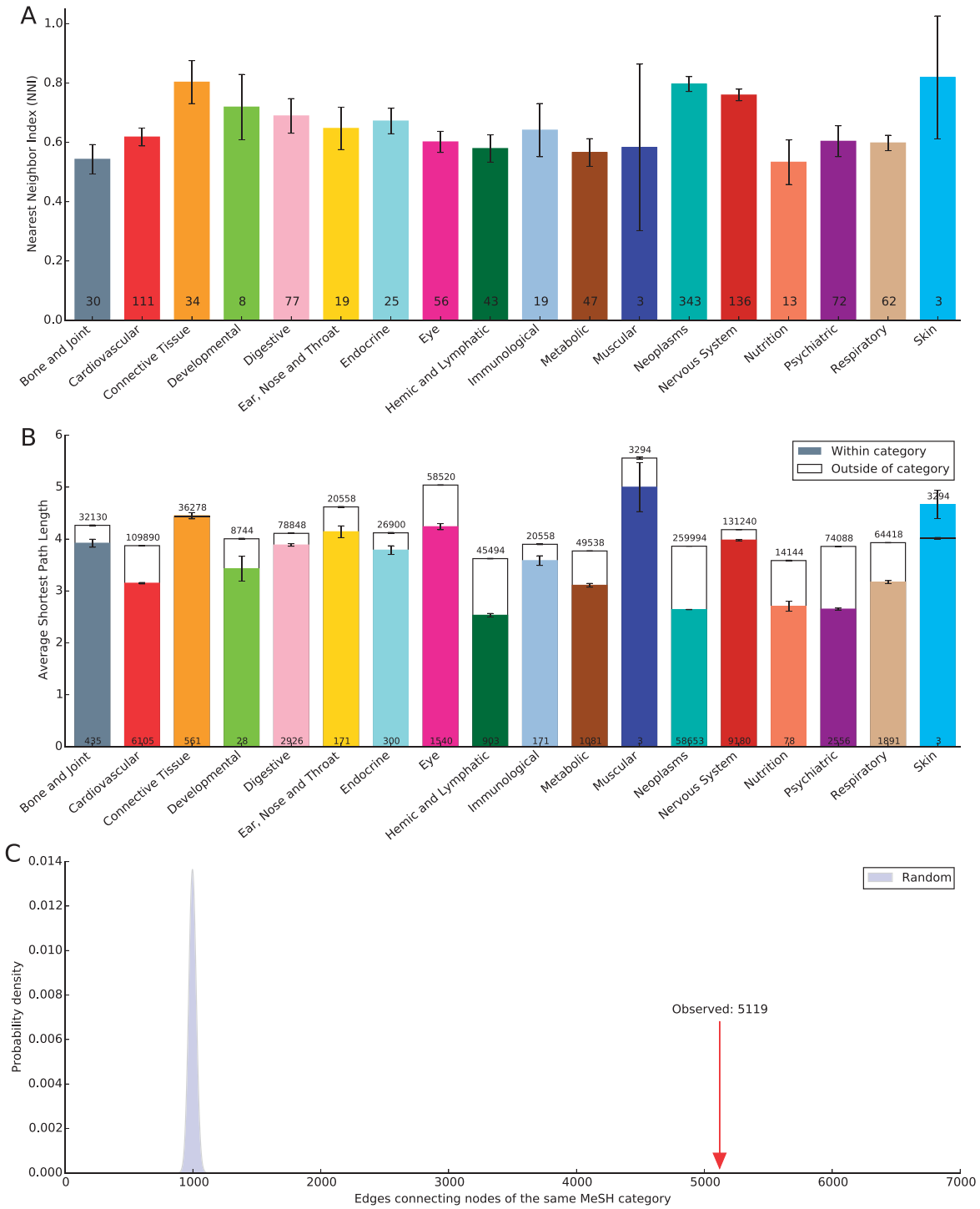

Figure 4. The disease-disease network (DDN) shows clustering by MeSH category. (**A**) The nearest neighbor index (NNI) for each MeSH disease category. Values significantly less than 1 indicate clustering in our visualization. Numbers on bars indicate how many diseases in the DDN are in each MeSH category. (**B**) The average shortest path length between nodes in the same category (solid bars) compared to the average shortest path between a node within a given category and all nodes outside the category (white bars). Solid bars that are significantly lower than white bars indicate more distinct clusters in the network. Error bars are the standard error. The numbers on the bottom of each bar indicate how many pairs of nodes are both within the category versus the numbers on top showing how many pairs of nodes have one within the category and one outside of the category. (**C**) The binomial distribution of edges between nodes of the same category if they were randomly placed on the graph (shaded area) compared to the observed number of edges in the disease-disease network (arrow).

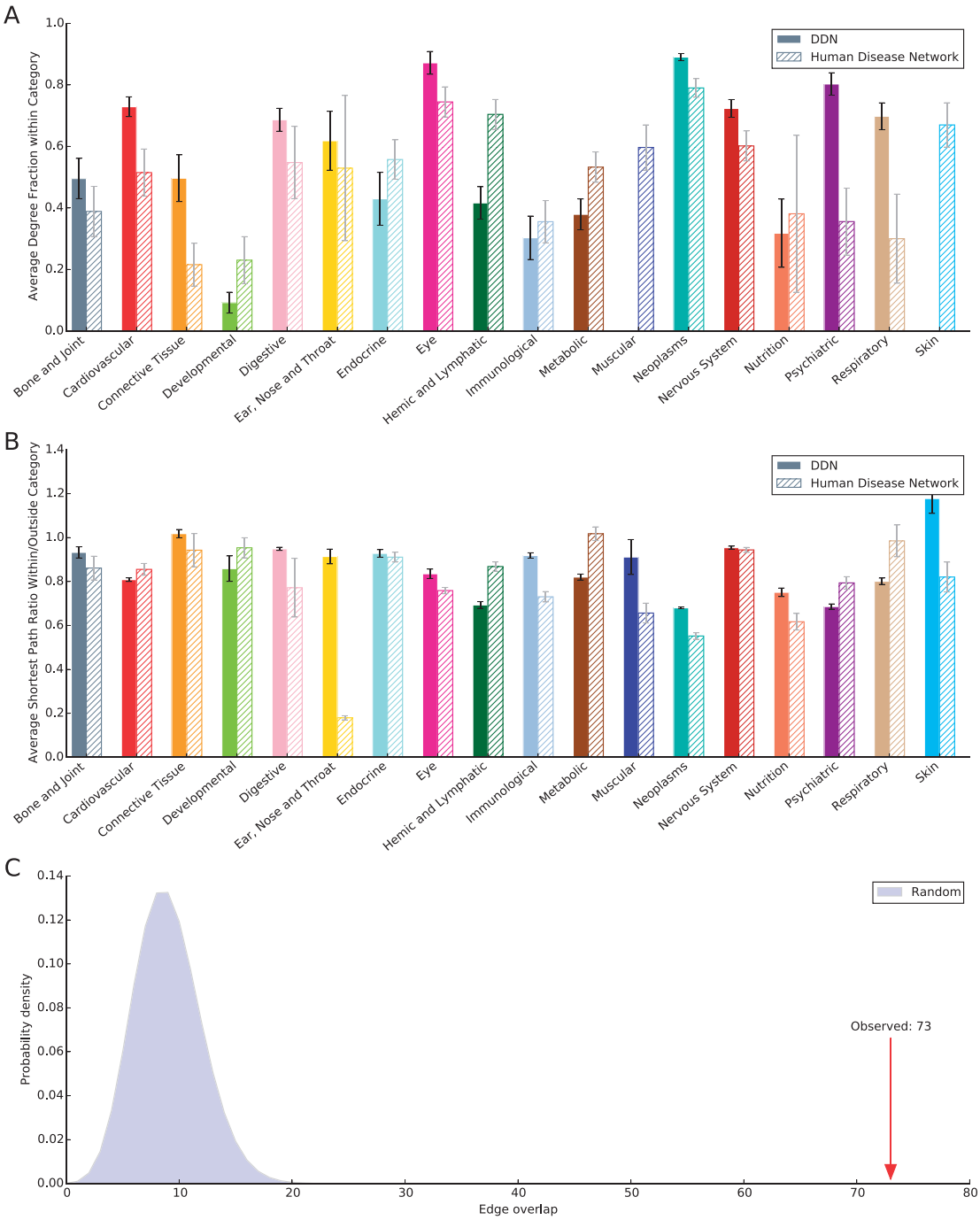

Figure 5. The disease-disease network (DDN) shows similarity in clustering of MeSH categories compared to the human disease network (HDN), which is based on a very different dataset. (**A**) The average degree fraction ratio for nodes within the same category. The ratio is the number of edges extending to nodes in the same category to the number of edges extending to nodes in different categories. The DDN is shown solid for each category, with the HDN shown in white with hatches next to it. In general, the DDN shows a similar or greater ratio than the HDN, demonstrating similar or even tighter clustering. (**B**) The average shortest path ratio for nodes within the same category. The shortest path ratio for a node is the ratio of the mean of the shortest path to every node in the same category to the mean of the shortest path to every node outside the category. As in (A), the solid bars correspond to the DDN and the white bars with hatches to the HDN. In general, the DDN shows a similar or smaller ratio than the HDN, demonstrating similar or possibly better clustering. (**C**) The binomial distribution of overlapping edges between the DDN and the HDN if edges were randomly placed between nodes, with the observed number shown by the arrow. The comparison is only made for nodes that are identical in the 2 graphs.

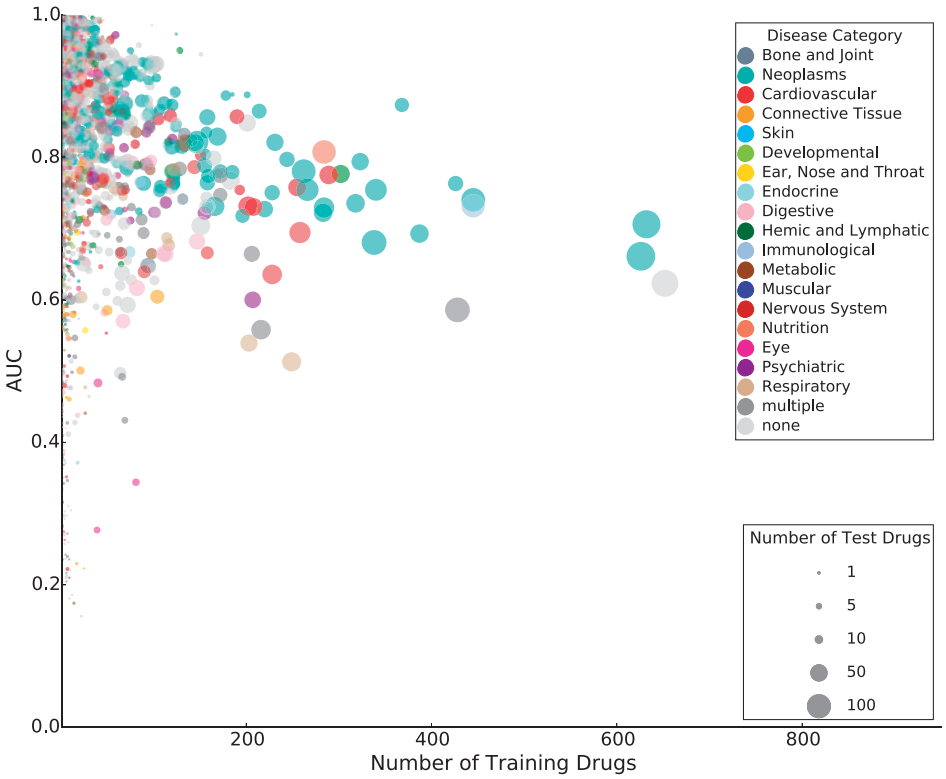

Figure 6. Area under the curve for drug predictions for individual diseases. The AUC for every disease is shown on the Y-axis, with the number of drugs in the training set on the X-axis and number of drugs in the test set for each disease as the size of each circle. The disease category is also shown as the color of the circle. Intuitively, more data should lead to more accurate predictions, but as the number of drugs in the dataset increases, the AUC actually decreases. This may be because diseases that have many drugs (more than 50–100) may be more prevalent and result in more random exploration of drugs than those with just a few.

literature supporting the connection between the disease and the predicted drug.

DISCUSSION

Our results demonstrate that clinical trial metadata can be used to infer disease relationships found in genetic data or medical taxonomies. Such agreement is surprising, given that the metadata used contains no explicit information about biology. The agreement suggests that our method succeeded in leveraging information latent in the collection of clinical trials to draw conclusions beyond what any single trial could reveal. The aggregate expert knowledge may reveal disease relationships a single group of experts may not have identified by themselves. This result opens the possibility that such latent information in clinical trial data or other types of clinical data can be used to uncover biological relationships that otherwise might only be found by using detailed biological data, by gaining access to large amounts of clinical data, or by conducting resource-intensive research. There are several promising avenues for use of the DDN as a resource to generate new hypotheses for biological and medical research.

The visualization of the DDN we presented provides a quick global reference of the therapeutic links among diseases conveying the underlying similarities among diseases. This similarity is based on the decision to run specific clinical trials as observed in ClinicalTrials.gov. The decision to conduct a trial is based on the summation of biological and medical knowledge, such as published research, proprietary in vitro or animal study results, clinical observations, results from previous clinical trials, and economic considerations. This represents a significant body of cross-disciplinary information leading to the decision to run any single trial. At present, lessons learned from this cross-discipline endeavor are shared in part through publications, reviews, and conference proceedings; collating this information for the entire body of clinical trials to derive lessons about human biology would be very time intensive. Instead, our representation of clinical trial metadata allows us to access that cross-discipline information implicitly to derive conclusions and lessons learned. The DDN we built demonstrates one way that access to such implicit or latent information can be used to draw conclusions beyond the information contained in any single trial, such as similarities between diseases.

As an example, asthma and inflammatory bowel diseases (Crohn's disease and ulcerative colitis) are closely related in the DDN map, even though the MeSH taxonomy classifies them differently and there is little direct connection between the diseases in the dataset. They are both inflammatory diseases, though, and recent research suggests that patients with asthma are at higher risk for inflammatory bowel disease.[38] Such similarities could stimulate hypotheses about related biological pathways, epidemiological connections, comorbidities in patients, or new indications for drugs.

As a computational tool, we showed that the DDN may be used to recommend drugs to test on a given disease. Similarly, for a given drug, one might predict which set of diseases would most likely benefit. Others have tried this approach by relating drugs using aggregated datasets as described earlier.[11–13]

The DDN may be most useful in combination with other data sources. For example, overlap of gene expression in 2 diseases in the HDN[7]

seemed to indicate that the genes might be involved in the same biochemical pathways in both diseases. The DDN also provides information about shared biochemical pathways, but from the perspective of drugs that could modulate those pathways. Together, the HDN and DDN could point more precisely to pathways or help distinguish between a genetic and environmental etiology. The DDN could be used in conjunction with datasets such as epigenetics, metabolomics, environmental factors, symptoms, and others that could be layered together to support inference.

In this work we limited ourselves to trial header information, based on our desire to focus on what we called superficial information and our intent to explore the limits of learning imposed by the availability of data. Our work shows that there is potential to extract more information from clinical trial data. Now that we have established this as a benchmark, future work could make use of additional information about each trial, such as inclusion/exclusion criteria or trial results. The thesaurus we built and the cleaned trial data will be extremely useful for expanding this work and may prove useful for other research in related areas such as meta-analysis.

For visualization and validation, we limited our exploration to a subset of the network. We filtered out some of the nodes for easier comparison with the HDN, such as infectious diseases that do not occur in the HDN. We filtered out edges for clearer visualization and to demonstrate the structure present in our data. Having validated a large subset of the network, future work could explore visualization methods on the unfiltered network, other filtering techniques to extract different meaning from the network, and the use of other similarity metrics and inferences on graphs.[39]

There is other information in the clinical trial data, such as the economic potential of drugs, special interest in orphan diseases, and the prevalence of diseases in developed countries. Such information may bias an attempt to draw inferences about biological relationships. In this work we focused on capturing all learning from trials. Future work using this learning for inference should account for bias, depending on the specific inference problem. One example where bias would not need to be removed would be in predicting which sets of diseases are least explored. Such predictions would be useful for determining what future trials would lead to the greatest increase in the understanding of diseases. The amount of learning indicated by the topology of the DDN does not always match a straightforward measure, such as the number of trials or drugs tested on a disease (Supplementary Figure 8).

Beyond the DDN as a resource, the approach we demonstrated may prove useful in other areas where latent information is contained in seemingly superficial data. Though it is anticipated that a flood of medical data will be released in the future, there is much more that can be done with the seemingly superficial data that is currently available. For example, "off-label" prescription data could be used in the same manner to uncover aggregated learning implicitly taking place by physicians in clinical practice.

Our results provide an example of using experimental data on humans, which is rare and valuable, to extract biologically useful information. This approach is different from the typical approach of learning biology mechanisms in models and then testing to see if they also hold in humans. Here we have shown how relationships can be derived from testing in humans and then explored to see if those relationships can improve understanding of biological mechanisms. As more clinical data does become available, it will be important to have tools like these in place to more rapidly uncover biological insights and discover effective treatments. For example, patient-level clinical trial data is becoming more available to researchers, but it is not clear how to compare 2 patients from 2 trials that were constructed for different purposes. As patient-level data becomes available, we see opportunities to extend this work to provide a structure for making such comparisons and posing research questions that do not depend on clinical endpoints.

CONCLUSIONS

We demonstrated that clinical trial metadata can be used to derive biologically meaningful disease relationships as tested using other disease networks and taxonomies. We therefore conclude that there is latent expert knowledge in the metadata. Our disease-disease network (DDN) shows a way to access that knowledge and to leverage the collective expert understanding of diseases. The relationships unique to the network can be used to generate new hypotheses for future biological and clinical research. This demonstrates a new strategy for leveraging research data on humans to advance our understanding of biological mechanisms. Furthering this approach to the translation of clinical data back to biological research will become even more important as more granular clinical data becomes available.

REFERENCES

1. Sung NS, Crowley WF, Genel M, *et al*. Central challenges facing the national clinical research enterprise. *JAMA*. 2003;289(10):1278–1287.
2. Adams CP, Brantner VV. Spending on new drug development. *Health Econ*. 2010;19:130–141.
3. De Angelis C, Drazen JM, Frizelle FA, *et al*. Clinical trial registration: a statement from the international committee of medical journal editors. *N Engl J Med*. 2004;351(12):1250–1251.
4. Drazen JM. Sharing individual patient data from clinical trials. *N Engl J Med*. 2015;372:201–202.
5. Zarin DA. Participant-level data and the new frontier in trial transparency. *N Engl J Med*. 2013;369(5):468–469.
6. 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.
7. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. The human disease network. *Proc Natl Acad Sci*. 2007;104(21):8685–8690.
8. Butte AJ, Kohane IS. Creation and implications of a phenome-genome network. *Nat Biotech*. 2006;24(1):55–62.
9. Zhou X, Menche J, Barabasi AL, Sharma A. Human symptoms-disease network. *Nat Commun*. 2014;5:4212.
10. Hidalgo CA, Blumm B, Barabasi AL, Christakis NA. A dynamic network approach for the study of human phenotypes. *PLoS Comput Biol*. 2009;5(4):e1000353.
11. Yildirim MA, Goh KI, Cusick ME, Barabasi AL, Vidal M. Drug-target network. *Nat Biotech*. 2007;25(10):1119–1126.
12. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. Drug target identification using side-effect similarity. *Science*. 2008;321(5886):263–266.
13. Tatnoetti NP, Ye PP, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. *Sci Transl Med*. 2012;4(125):125–131.
14. Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a newtork-based approach to human disease. *Nat. Rev Genetics*. 2011;12:56–68.
15. Fox, RJ, Miller DH, Phillips T, *et al*. Placebo-controlled phase 3 study of oral bg-12 or glatiramer in multiple sclerosis. *N Engl J Med*. 2012;367(12):1087–1097.
16. van 't Veer LJ, Dai H, van de Vijver MJ, *et al*. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 2002;415(6871):530–536.

17. Klebe G, Abraham U, Mietzner T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J Med Chem*. 1994;37(24):4130–4146.
18. Wermuth CG. Similarity in drugs: reflections on analogue design. *Drug Discovery Today*. 2006;11(7-8):348–354.
19. Vilar S, Uriarte E, *et al*. Similarity-based modeling in large-scale prediction of drug-drug interactions. *Nat. Protocols*. 2014;9(9):2147–2163.
20. Keiser MJ, Setola V, Irwin JJ, *et al*. Predicting new molecular targets for known drugs. *Nature*. 2009; 462(7270):175–181.
21. ClinicalTrials.gov Web site. See Drug Interventions by First Letter. https://clinicaltrials.gov/ct2/search/browse?brwse=intr_alpha_all. Accessed December 16, 2015.
22. Tasneem A, Aberle L, Ananth H, *et al*. The database for aggregate analysis of clinical trials (AACT) and subsequent regrouping by clinical specialty. *PLoS One*. 2012;7(3):e33677.
23. Califf RM, Zarin DA, Kramer JM, *et al*. Characteristics of clinical trials registered in ClinicalTrials.gov, 2007-2010. *JAMA*. 2012;307(17): 1838–1847.
24. Hirsch BR, Califf RM, Cheng SK, *et al*. Characteristics of oncology clinical trials. *JAMA Int Med*. 2013;173(11):972–979.
25. Inrig, JK, Califf RM, Tasneem A, *et al*. The landscape of clinical trials in nephrology: a systematic review of ClinicalTrials.gov. *Am J Kidney Dis*. 2014;63(5):771–780.
26. Wu M, Sirota M, Butte AJ, Chen B. Characteristics of drug combination therapy in oncology by analyzing clinical trial data on ClincialTrials.gov. *Pac Symp Biocomput*. 2015;68–79.
27. He Z, Carini S, Hao T, , Sim I, Weng C. A method for analyzing commonalities in clinical trial target populations. *AMIA Annu Symp Proc*. 2014;1777–1786.
28. Hao T, Rusanov A, Boland MR, Weng C. Clustering clinical trials with similar eligibility criteria features. *J Biomed Inform*. 2014;52:112–120.
29. Lipscomb CE. Medical Subject Headings (MeSH). *Bull Med Library Assoc*. 2000;88(3):265–266.
30. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*. 2004;32:D267–D270.
31. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc*. 2010;17:229–236.
32. Zarin DA, Keselman A. Registering a clinical trial in ClinicalTrials.gov. *CHEST*. 2007;131(3):909–912.
33. Barabasi AL, Albert R. Emergence of scaling in random networks. *Science*. 1999;286(5439):509–512.
34. Clark PJ, Evans FC. Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology*. 1954;35(4):445–453.
35. Schafer JB, Frankowski D, Herlocker J, Sen S. Collaborative filtering recommender systems. In: P Brusilovsky, A Kobsa, W Nejdl, eds. *Lecture Notes in Computer Science*, Vol. 4321. Berlin: Springer; 2007:291–324.
36. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recogn*. 1997;30(7):1145–1159.
37. North BV, Curtis D, Sham PC. A note on the calculation of empirical P values from Monte Carlo procedures. *Am J Hum Genet*. 2002;71(2): 439–441.
38. Brassard P, Vutcovici M, Ernst P, *et al*. Increased incidence of inflammatory bowel disease in Québec residents with airway diseases. *Eur Respir J*. 2015;45(4):962–968.
39. Murphy DP, Weiss Y, Jordan MI. Loopy belief propagation for approximate inference: an empirical study. In: *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*. San Francisco: Morgan Kaufmann Publishers; 1999:467–475.
40. Huser V, Cimino JJ. Evaluating adherence to the International Committee of Medical Journal Editors' policy of mandatory, timely clinical trial registration. *J Am Med Inform Assoc*. 2013;20:e169–e174.
41. Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. *J Soviet Physics Doklady*. 1966;10(8):707–709.
42. Ratclif JW, Metzener D. Pattern matching: the Gestalt approach. *Dr Dobb's J*. 1988;13(7):46.
43. Kobourov SG. Force-directed drawing algorithms. In: R Tamassia, ed. *Handbook of Graph Drawing and Visualization*. Boca Raton, FL: CRC Press; 2014:383–408.
44. Fruchterman TMJ, Reingold EM. Graph drawing by force-directed placement. *Softw: Pract Exper*. 1991;21(11):1129–1164.
45. Kamada T, Kawai S. An algorithm for drawing general undirected graphs. *Inform Process Lett*. 1989;31(1):7–15.
46. Eades P. A heuristic for graph drawing. *Congressus Numerantium*. 1984;42:149–160.
47. Jacomy M, Venturini T, Heymann S, Bastian M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS ONE*. 2014;9(6):e98679.
48. Kirkpatrick S, Gelatt CD, Vecchi MP. Optimdrawing by simulated annealing. *Science*. 1983;220(4598):671–680.
49. Mills JD, Hadley K, Bailes JE. Dietary supplementation with the omega-3 fatty acid docosahexaenoic acid in traumatic brain injury. *Neurosurgery*. 2011;68(2):474–481.
50. Inoue M, Arakawa A, Yamane S, Kadonosono K. Short-term efficacy of intravitreal aflibercept in treatment-naive patients with polypoidal choroidal vasculopathy. *Retina*. 2014;34(11):2178–2184.
51. Tunks RD, Clowse MEB, Miller SG, *et al*. Maternal autoantibody levels in congenital heart block and potential prophylaxis with antiinflammatory agents. *Am J Obstetrics Gynecol*. 2013;208(1):64–67.
52. McElroy SL, Guerdjikova AI, Mori N, O'Melia AM. Pharmacological management of binge eating disorder: current and emerging treatment options. *Therapeutics Clin Risk Manag*. 2012;8:219.
53. Hosono K, Endo H, Takahashi K, *et al*. Metformin suppresses colorectal aberrant crypt foci in a short-term clinical trial. *Cancer Prevent Res*. 2010;3(9): 1077–1083.
54. Yang SI, Chung WJ, Jung SH, Choi DY. Effects of inhaled iloprost on congenital heart disease with Eisenmenger syndrome. *Pediatric Cardiol*. 2012;33(5):744–748.
55. Pesek R, Fox R. Successful treatment of Schnitzler syndrome with canakinumab. *Cutis*. 2014;94(3):E11–E12.
56. Vanderschueren S, Kockaert D. Canakinumab in Schnitzler syndrome. *Semin Arthritis Rheum*. 2013;42(4):413–416.