Reforming disease classification system—are we there yet?

Mikhail G. Dozmorov

Department of Biostatistics, Virginia Commonwealth University, Richmond, Virginia, USA

Correspondence to: Mikhail G. Dozmorov. Department of Biostatistics, Virginia Commonwealth University, 830 East Main Street, Richmond, Virginia 23298, USA. Email: mikhail.dozmorov@vcuhealth.org.

Comment on: Zhou X, Lei L, Liu J, *et al.* A Systems Approach to Refine Disease Taxonomy by Integrating Phenotypic and Molecular Networks. EBioMedicine 2018;31:79-91.

Submitted Sep 04, 2018. Accepted for publication Sep 12, 2018.
doi: 10.21037/atm.2018.09.36
View this article at: http://dx.doi.org/10.21037/atm.2018.09.36

What constitutes a disease? How similar is one disease to another? How is a disease of one patient similar to a disease of another patient? These questions are probably as old as the medical practice itself. Answering them would allow for better understanding of underlying mechanisms, guide diagnosis, improve prognosis, justify drug repurposing.

Clinicians frequently observe that symptoms or certain patients diagnosed with different diseases co-occur more frequently than can be expected by chance, suggesting common mechanisms. These observations gave rise to the constitutional principle of the current view of disease classification that relates diseases to each other by shared clinical signs, pathophysiology, etiology, or cellular endophenotypes. Several disease taxonomies have been created, such as the Human Phenotype Ontology (HPO) that structures all phenotypic abnormalities that are commonly encountered in human monogenic diseases as a directed acyclic graph (DAG) (1), the Disease Ontology (DO) that uses semantic similarity measures to integrate several vocabularies of medical terms into a DAG of disease hierarchy (2). The International Classification of Diseases (ICD) is the most widely used clinically-oriented hierarchical vocabulary of diagnostic codes. Maintained by the World Health Organization, it is extensively used as a health care classification system in the US (ICD-9 edition) and in Europe, Canada and Australia (ICD-10 edition). Currently, the ICD-9 taxonomy remains the most widespread and accessible as a *de facto* standard of disease similarities.

Despite its wide popularity, the ICD-9 classification system lags behind modern disease research by not exploiting the rapid growth of our understanding of the molecular mechanisms of disease. This may be best characterized by the fact that many diseases in the current disease taxonomies have high genetic heterogeneity or manifestation diversity (3). Although the 11th edition of the International Classification of Diseases (ICD-11), expected to be more comprehensive than its predecessors, has been released in 2018, it still does not include molecular properties of diseases. Furthermore, its adoption will begin in 2022. This lack of depth in disease representation limits the opportunities for tailoring treatment to a patient's pathophysiology. Thus, the reform of ICD-based disease taxonomy is warranted.

Efforts to devise a better disease classification system based on molecular insights have blossomed with the discovery of the genetic code. The initial search for the genetic component of the diseases has uncovered over 1,000 phenotypes associated with single genetic changes, leading to the creation of Online Mendelian Inheritance of Man (OMIM) database (4). However, further research uncovered a more complex picture in which perturbation of multiple molecular mechanisms and environmental factors contribute to disease manifestation. Consequently, disease similarity can be represented as a network built on shared molecular features of disease-associated genes, proteins, metabolites, etc. (5,6). Nodes in such networks are typically represented by genes, while edges correspond to some functional relationships (e.g., co-expression, protein-protein interactions). Disease similarity within such networks can be explored using classical network metrics, such as community detection, betweenness centrality, etc. The intuitive expectation is that diseases sharing similar molecular mechanisms will form coherent modules with similar genes, gene expression profiles, have shared genomic variants, higher PPI interactions, higher co-morbidity, share common pathways and gene ontologies (7). Numerous researches showed that this intuition is viable and utilizing any single molecular data type provides novel insights into disease similarity, reviewed in (8). However, disease classification built using any single data type, although conceptually simple, uses only a fraction of molecular information thus limiting disease classification efforts.

Approaches that integrate multiple sources of molecular interactions give a more holistic view of disease similarities that contains more information than the sum of its parts (6,9,10). Perhaps the most natural way of combining omics data is integrating gene information (co-expression, genomic variant and the associated gene overlap, healthy-disease differential expression) with protein-protein interactions, and use the resulting network to maximize disease similarity search (11). Known gene interaction data, such as canonical pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome, functional relationships from the Gene Ontology (GO) database (12,13), as well as text-based (aka semantic similarity) measures (14) were also utilized for refining disease similarity networks. Data integration methods that utilize multiple types of data, including omics, ontological, and textual data, established a strong foundation for integrative approaches in the search for disease similarities, reviewed in (8).

The benefits of integrative approaches were recently demonstrated by Zhou *et al.* who proposed to redefine the modern disease classification system, ICD-9, by defining a new system for disease classification, referred to as new classification of diseases NCD (NCD) (15). It is developed by integrating the curated tree of 1,883 ICD disease codes, grouped into chapters, with phenotype similarity measures, shared genes and protein-protein interaction modules. A metric to evaluate the quality of ICD disease taxonomy convincingly showed that the grouping of ICD chapters does not agree with the natural topological groupings of the corresponding molecular networks. Consequently, a novel algorithm was developed to generate the possible associated additional disease categories for a given disease with the corresponding molecular association scores. This algorithm was used to define the NCD. The diseases in NCD were grouped into a network containing 17 new disease chapters (NC) and 223 subcategories that have significantly higher network modularity than the original ICD chapters. Furthermore, the phenotypic and molecular links between the diseases in an NC are much denser compared to the ICD taxonomy. Benchmarking of NCD showed that it better captures disease similarity in terms of gene similarity, gene ontology and phenotypic similarity. In summary, NCD represents a refinement of ICD-9 disease taxonomy by capturing the molecular diversity of diseases and defining clearer boundaries in terms of both phenotypic similarity and molecular associations.

One important aspect of publishing large-scale disease classification studies is data availability. Many studies implement disease similarity search as web-based tools or programming packages and provide disease similarity data for download, reviewed in (8). Although the work of Zhou *et al.* (15) does not offer a software solution for disease classification, it is a treasure trove of data for researchers interested in disease similarity research. Given the significant efforts in data collection, curation, and integration of disparate data sources, the availability of data used at all stages of the analysis performed by Zhou *et al.* will help to ensure reproducibility and provide an essential resource for disease similarity researches.

Does the New Disease Classification give us the final disease classification system? Although we are getting better at understanding disease relationships, the answer is no. One reason is that the NCD system is a better reclassification of 1,883 ICD-9 codes, while many codes could not be mapped to genes and/or phenotypes. Furthermore, many ICD-9 codes may belong to multiple categories and subcategories in the NDC taxonomy (~40% in Zhou *et al.* work), creating ambiguity in disease definition. This multi-classification problem is particularly pronounced in cancer and infectious diseases that have diverse molecular network mechanisms and tend to interact with diseases from different chapters. Thus, a more precise vocabulary of disease codes is needed for a better disease classification taxonomy. Although the ICD-10 and ICD-11 editions have more disease codes, they have a similar structure to the currently used ICD-9 edition. Utilizing a more diverse disease coding system coupled with integrative analyses of the underlying molecular mechanisms will bring us a step closer to the better disease classification system.

Disease classification research is incomplete without considering the three-dimensional (3D) structure of the genome (16). The genome is organized into complex higher-order structures by folding of the DNA into coiled chromatin fibers, chromosome domains, and ultimately chromosomes. These structures are non-random, with active euchromatin and inactive heterochromatin occupying separate environments. Chromosomes themselves occupy distinct territories, and further subdivided into a hierarchy of topologically associated domains (TADs) and, on the most local level, chromatin loops (17). Genomic variants, such as copy number variations (CNVs) and single nucleotide polymorphisms (SNPs), all have been shown to disrupt chromatin interactions that mark TAD boundaries. These disruptions lead to gene expression changes and disease manifestation (18). Changes in chromatin interactions are only now emerging as a hallmark of cancer (19) and other diseases. Thus, disease similarity metrics may be complemented by the similarity in the 3D genomic structures, and the location of genomic variants within them, providing a more holistic understanding of similarities and differences among diseases.

As diseases are products of complex gene and environmental interactions (6), disease classification systems can be further improved by considering the environmental effect. Comorbidity measures represent an indirect way to measure the effect of the environment on disease manifestation. Increased comorbidity between diseases is frequently used as a confirmatory and/or discovery step in understanding disease similarity (20). Electronic Health Records (EHRs) represent a large corpus of data about disease comorbidities. Importantly, although EHRs do not explicitly contain information about the underlying molecular mechanisms, they record real-life manifestation of them, thus providing a complementary metric for measuring disease similarity (21). Consequently, their integration with genetic (22) and PPI (23) networks have been shown to augment our understanding of the molecular mechanisms of diseases (24).

Search for disease similarity continues with the growing amount of omics data and EHRs, and with the concurrent development of machine and deep learning approaches that effectively learn various aspects of disease similarity from these big data. For example, artificial neural networks have been successfully applied to large collections of EHRs for predicting disease-associated genes, classify patients, and predict future medical outcomes (8,24). Machine and deep learning approaches integrating multiple data sources show promise in providing a maximally accurate disease classification system. All the aforementioned considerations need to be integrated for the maximally comprehensive classification system of human diseases.

Acknowledgements

None.

Footnote

Conflicts of Interest: The author has no conflicts of interest to declare.

References

1. Robinson PN, Köhler S, Bauer S, et al. The human phenotype ontology: A tool for annotating and analyzing human hereditary disease. Am J Hum Genet 2008;83:610-5.
2. Osborne JD, Flatow J, Holko M, et al. Annotating the human genome with disease ontology. BMC Genomics 2009;10 Suppl 1:S6.
3. McClellan J, King MC. Genetic heterogeneity in human disease. Cell 2010;141:210-7.
4. McKusick VA. Mendelian inheritance in man and its online version, omim. Am J Hum Genet 2007;80:588-604.
5. Barabási AL, Oltvai ZN. Network biology: Understanding the cell's functional organization. Nat Rev Genet 2004;5:101-13.
6. Schadt EE. Molecular networks as sensors and drivers of common human diseases. Nature 2009;461:218-23.
7. Menche J, Sharma A, Kitsak M, et al. Disease networks. uncovering disease-disease relationships through the incomplete interactome. Science 2015;347:1257601.
8. Dozmorov MG. Disease classification: from phenotypic similarity to integrative genomics and beyond. Brief Bioinform 2018. [Epub ahead of print].
9. McCarthy MI, Smedley D, Hide W. New methods for finding disease-susceptibility genes: Impact and potential. Genome Biol 2003;4:119.
10. Krishnan A, Taroni JN, Greene CS. Integrative networks illuminate biological factors underlying gene-Disease associations. Current Genetic Medicine Reports 2016;4:155-62
11. Suthram S, Dudley JT, Chiang AP, et al. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. PLoS Comput Biol 2010;6:e1000662.
12. Franke L, van Bakel H, Fokkens L, et al. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am J Hum Genet 2006;78:1011-25.
13. Aerts S, Lambrechts D, Maity S, et al. Gene prioritization through genomic data fusion. Nat Biotechnol 2006;24:537-44.
14. Mathur S, Dinakarpandian D. Finding disease similarity

based on implicit semantic similarity. J Biomed Inform 2012;45:363-71.

15. Zhou X, Lei L, Liu J, et al. A Systems Approach to Refine Disease Taxonomy by Integrating Phenotypic and Molecular Networks. EBioMedicine 2018;31:79-91.
16. Babu D, Fullwood MJ. 3D genome organization in health and disease: Emerging opportunities in cancer translational medicine. Nucleus 2015;6:382-93.
17. Denker A, de Laat W. The second decade of 3C technologies: Detailed insights into nuclear organization. Genes Dev 2016;30:1357-82.
18. Spielmann M, Lupiáñez DG, Mundlos S. Structural variation in the 3D genome. Nat Rev Genet 2018;19:453-67.
19. Valton AL, Dekker J. TAD disruption as oncogenic driver. Curr Opin Genet Dev 2016;36:34-40.
20. Melamed RD, Emmett KJ, Madubata C, et al. Genetic similarity between cancers and comorbid mendelian diseases identifies candidate driver genes. Nat Commun 2015;6:7033.
21. Roque FS, Jensen PB, Schmock H, et al. Using electronic patient records to discover disease correlations and stratify patient cohorts. PLoS Comput Biol 2011;7:e1002141.
22. Davis DA, Chawla NV. Exploring and exploiting disease interactions from multi-relational gene and phenotype networks. PLoS One 2011;6:e22670.
23. Wang Q, Liu W, Ning S, et al. Community of protein complexes impacts disease association. Eur J Hum Genet 2012;20:1162-7.
24. Gligorijevic D, Stojanovic J, Djuric N, et al. Large-scale discovery of disease-disease and disease-gene associations. Sci Rep 2016;6:32404.

Cite this article as: Dozmorov MG. Reforming disease classification system—are we there yet? Ann Transl Med 2018;6(Suppl 1):S30. doi: 10.21037/atm.2018.09.36