Uncovering Small Secrets in Big Data Sets
How Math Can Identify Biology in Rare Conditions
(Pediatric Pulmonary Hypertension)

Mardi Gomberg-Maitland, Rogerio Souza

Complex patterns recognition is an evolutionary capability that differentiates humans from any other animal. Whereas environment recognition and emotional state changes are seen in many different mammals, the cognitive repertoire of humans is more comprehensive and multifaceted. Historically, progressive development in medicine has been somehow linked to this particular ability. Clinical syndromes, diagnostic tools, and treatments have been developed based on classification systems. Classification systems have also been incorporated in the clinical practice by our ability to group similar patients according to symptoms, pathological findings, and pathophysiological aspects of any determined condition. Classification systems are dynamic tools that incorporate advancing technologies and allow better communication among physicians and researchers in prediction of treatment response and prognosis.

Because biotechnology has improved, our ability to phenotype disease with genomics, proteomics, and imaging modalities is now sophisticated. Usual statistical analysis lacks the discriminative ability to handle so many factors. Advanced statistical modeling and artificial intelligence are the next step in advancing medical science to handle all the data and lead to a more scientific phenotyping of disease. One such method is neuronal networks. Like the brain, many factors (nodes) are interlinked in multiple circuits to form a network. This allows for multiple interactions to occur with multiple linkage combinations. Networks provide information on biological processes and clinical phenotypic differences between patients with different demographics, disease progression rates, and mortalities.

Pulmonary hypertension (PH) is a heterogeneous disease with its current classification schema based on commonalities in cause, pathology, and expert opinion. PH is currently classified into 5 groups: group 1 pulmonary arterial hypertension, group 2 PH associated with left heart disease, group 3 PH associated with hypoxic lung disease, group 4 PH associated with chronic thromboembolism, and group 5 PH associated with multifactorial mechanisms. Pulmonary arterial hypertension remains rare and fatal despite the advent of better diagnostic tools, algorithms, and treatments. The average prognosis without therapy is 2.8 years, and despite educational efforts, patients are diagnosed late in their disease state. Surprisingly, despite geographic differences, much of the data are similar.

Nevertheless, our knowledge base on pediatric PH is significantly more narrow. There are far fewer registries incorporating pediatric patients and few clinical trials including pediatric patients. Similar to adult observational registry and cohort studies, pediatric studies are subject to survival and referral bias. The recognition that childhood onset PH may have unique pathogeneses not observed in adults prompted the development of a new taxonomy. The Panama classification highlighted more common features seen in pediatric onset disease, including fetal and developmental origins of vascular disease. However, the current PH classification tried to acknowledge these pediatric specificities separately to keep the same system for children and adults, facilitating the transition of care from one phase (pediatric) to the next (adulthood). Both systems have clear limitations, most of them based on the paucity of data available on pediatric PH. The recent center-based registry TOPP study (Tracking Outcome and Practice in Pediatric PH) is the largest pediatric PH registry to date including 31 centers in 19 countries. TOPP has helped with our understanding of pediatric PH demographics and care despite its inability to identify precise phenotypes or novel biological pathophysiologic mechanism.

In this issue, Ong et al. aimed to use network statistics to improve clinical definitions and identify clinically relevant patterns of disease in pediatric PH patients with hope that this would lead to earlier and improved diagnostic specificity. This is the first study in PH using machine learning techniques to improve our understanding of rare disease. Using an Aetna Incorporation claims data set of 6,943,263 children from January 2010 to May 2013, they enrolled children with ≥2 visits associated with PH whom met the criteria for PH (1583 [0.02%]). Initially, this entailed an echocardiographic diagnosis, but they also ran their statistical methodology to only include patients who also had cardiac catheterization, the gold standard diagnostic tool. Acknowledging the limits of the United States International Classification of Diseases, Ninth Revision codes for disease and the coding process itself, the data set is robust in numbers.

The details of the mathematical modeling are important to understand for the nonstatistical expert. It is the first time that investigators in PH have used a more organized dual modeling
approach. The authors performed usual analyses with comparative statistics and then did a Bayesian comorbidity network. A Bayesian network allowed the authors to model 186 comorbidities found in children with PH versus those without PH. This type of analysis defined dependent and independent relationships, the presence of a comorbidity either increased or decreased the likelihood of the other occurring; the authors only included dependent positive relationships. This is not a new statistical concept. What is novel is the capability to use this technique in clinical research because technology is now markedly advanced.

The network development to best define the relationships within the existing data is a complex series of statistical steps. The authors appropriately used a noisy-OR model so that the influence of each parent on a node was independent of other parents. They assigned weights by calculating the conditional probability of each node. To prevent overfitting, they applied boot strap modeling. This allowed them to validate their findings because bootstrapping takes multiple random samples to create random models from their data set. They also used known clinical differences in the pathogeneses of PH to help them form their networks. The third method used to limit noise and improve diagnostic accuracy was to only use comorbidities that were significantly associated with PH in bivariate analyses compared with the general population. The authors then applied network clustering to create highly associated networks, random walk clustering. Identified networks were then re-evaluated by clinicians and 1 researcher. They evaluated the inter-rater reliability and resolved discrepancies by statistical analyses and consensus. Finally, they performed a literature search on evidence to support and determine the clinical reliability of their findings.

The authors’ network successfully validated existing classification systems of well-established subtypes in addition to identifying rare subtypes. They acknowledged that International Classification of Disease codes often did not differentiate group 2 PH (associated with left heart disease) well. Some of this is inherent in the International Classification of Disease coding’s inability to accurately capture complex congenital heart disease. The database used is a US-based insurance data set, and thus, the numbers of patients with certain pathogeneses, such as human immunodeficiency virus and schistosomiasis, were low or nonexistent. They also acknowledged that despite previous research suggesting that diagnosis based on ≥2 encounters is valid, perhaps using a more stringent definition would change their findings. However, their sensitivity analysis considering only diagnoses with ≥4 encounter visits remained significant.

Importantly, they found several subtypes documented in only a few case studies for which systematic association remains lacking. Modeling identified subtypes consistent with several well-described genetic syndromes not previously examined for PH. (see Table 4 of article) Network-derived comorbidity clusters included juvenile idiopathic arthritis, hemophagocytic syndrome, glycogen storage disease, hereditary muscular dystrophy, cardiomyopathy, adrenogenital disorders, pulmonary collapse, nutritional deficiencies, Cri-du-chat syndrome, situs inversus, congenital spleen anomaly, and Turner and Prader-Willi syndromes. Although the recognition of these associations does not establish the potential pathophysiological mechanisms that might be associated with the development of PH, or perhaps already encompassed by the current classification systems, the authors acknowledge that their approach certainly brings focus to less recognized or more rare conditions that would not otherwise be associated with PH.

It is clear that combining advanced methodology with machine learning, clinical acumen, data, and observation has the potential to elucidate new opportunities in understanding disease. This is certainly relevant in less frequent conditions for which prospective evaluation represents a bigger challenge but no less important in common diseases. The use of networks in globally represented data sets creates potential for new phenotypes to help accelerate discovery and prompt new avenues of research in pathogenesis, susceptibility, and perhaps longevity. Clinical scientists need to pair with statistical and information technology experts to expedite this process. We commend the authors and hope that this type of research will continue in both common and less common diseases.

Disclosures

None.

References

11. Ong M-S, Mullen MP, Austin ED, Szolovits P, Natter MD, Geva A, Cai T, Kong SW, Mandl KD. Learning a comorbidity-driven taxonomy of


Key Words: Editorials ■ articial intelligence ■ heart diseases ■ hypertension, pulmonary ■ statistics
Uncovering Small Secrets in Big Data Sets: How Math Can Identify Biology in Rare Conditions (Pediatric Pulmonary Hypertension)

Mardi Gomberg-Maitland and Rogerio Souza

Circ Res. 2017;121:317-319
doi: 10.1161/CIRCRESAHA.117.311511

Circulation Research is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2017 American Heart Association, Inc. All rights reserved.
Print ISSN: 0009-7330. Online ISSN: 1524-4571

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circres.ahajournals.org/content/121/4/317

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation Research can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Circulation Research is online at:
http://circres.ahajournals.org/subscriptions/