Editorial

Quantitative Models for Causal Analysis in the Era of Genome Wide Association Studies

Causal inference in health research is a complex endeavor partly because the biomedical enterprise involves researchers from many disciplines including clinical medicine, epidemiology, genetics, basic sciences such as pathology and cell biology, and the behavioral sciences. A multidisciplinary approach is often needed to study health concerns and interpret findings, drawing upon expertise from epidemiologists, statisticians, physicians, nurses, geneticists, psychologists, and other practicing clinicians and researchers. In addition to the diversity of scientific disciplines and professions that are represented in many study groups, the range of health topics that can be studied is large and can include physical injuries such as traumatic brain injury; pain syndromes and other neurological conditions; chronic health conditions such as obesity, cancer, respiratory illnesses, and cardiovascular disease, gastrointestinal illnesses such as irritable bowel syndrome, infectious diseases such as H1N1 influenza and hepatitis C, psychiatric conditions such as post traumatic stress syndrome, depression, and suicide, adverse reproductive outcomes, and other health problems and concerns. Another feature of health research is that a range of study designs are employed by researchers including surveillance systems, observational studies with a case-control or cohort design, cross-sectional surveys, and randomized controlled trials. In recent years, observational studies include the large platforms of cases and controls that are identified for genome-wide association studies [1, 2]. In addition to statistical geneticists, the researchers who analyze data from genome-wide association studies and proteomics research often include persons with expertise in bioinformatics or machine learning techniques.

These three features of health research (diversity of scientific disciplines, wide variety of health topics of interest, and alternative study designs) create both challenges and opportunities for researchers attempting to identify causal associations with possible etiologic agents and new therapeutic targets, so that research findings can be translated into targeted clinical interventions and evidence-based therapies. For example, in studies with an observational design, where assignment of exposures is not under control of the investigators, assessments of causality can be more challenging than in randomized trials [3, 4].

Investigations into the distribution and determinants of health conditions attempt to gain new knowledge through observation and inductive logic. Causal criteria commonly cited in epidemiology include temporal order of exposure and disease, biologic gradient or dose-response curve, biologic plausibility, biologic coherence, and consistency of findings, although some authors have recommended subsets of the criteria or refined definitions [5-7]. The strength of the observed association is also important in some assessments of causality. Criteria for causal criteria are widely used as a heuristic aid for assessing whether associations observed in epidemiologic research are causal although criteria-based methods provide only general guidelines for assessing the causality of associations rather than a strict checklist for identifying a causal relationship [3, 4]. The model of sufficient component causes [8] is widely used in epidemiology as a framework for teaching and understanding multicausality. A sufficient component cause is made up of a number of components, no one of which is sufficient for the disease or adverse health condition on its own [4, 8]. Diseases and adverse health conditions can be caused by more than one causal mechanism and each causal mechanism involves the combined action of several component causes. For example, both genetic factors and environmental exposures may have a role in the development of neurologic conditions such as amyotrophic lateral sclerosis. Other examples of diseases caused by interactions between genes and environment include complex, common diseases such as cancer, coronary heart disease, and diabetes [1].

A large and growing literature has dealt with statistical modeling approaches for estimating causal parameters or identifying causal associations using data from observational studies [9-13]. However, much of this important literature has not dealt directly with the special challenges that arise in causal assessments of data from genome-wide association studies including information about environmental exposures. Recent advances in genetics have challenged traditional frameworks for causal inference in observational research [14].

The goal of this article is to consider challenges that arise in causal assessments of data from genome-wide association studies, which utilize high throughput genotyping technologies to analyze biological specimens collected from large numbers of cases and controls for up to one million single nucleotide polymorphisms (SNPs) [1, 2]. Before considering those challenges, I briefly discuss key developments in quantitative models for causal analysis: counterfactual analysis and graphical causal models and structural equations modeling. I then provide a summary of quantitative techniques for analyzing data from genome-wide association studies and related gene expression and proteomic data, and offer some recommendations for causal assessments of results from such studies.

COUNTERFACTUAL MODELS

Several authors have described how the counterfactual model for causal analysis or potential-outcomes model can be applied to observational data from health studies [11, 12, 15, 16]. Counterfactual models specify what would happen under alternative possible patterns of exposure and provide a basis for quantitative analyses of exposure effects [11]. Statistical approaches to causal analysis require that the exposure of interest (X) and the outcome Y be measurable quantities [16].

Suppose that i = 1… N represents a population of persons under study. As Greenland and Brumback [11] explained, the counterfactual model of causation assumes that: 1) each individual could have received any one of the exposure levels, and 2) for each individual i and exposure level X_j, at the time of exposure (or at random assignment of treatment), the outcome that individual i would have if the individual gets exposure level X_j exists, even if the individual does not actually get X_j. This value is referred to as the potential outcome of individual i

1874-9240/10 2010 Bentham Open

under exposure level X_j. The counterfactual model treats the potential outcome or outcomes as if they were baseline covariates or fixed from the start of follow-up [17].

The second assumption noted above can be restated in the following way: for each individual i and each exposure level $X_{j,}$ a potential-outcome variable Yij can also be defined that represents the outcome of the individual under that exposure [11]. A further assumption in many counterfactual modeling applications is that the potential outcomes of each individual are independent of the exposures (or treatment) and outcomes of other individuals.

Yij is the indicator of the actual outcome for individual i if individual i has exposure level X_j. Otherwise, Yij may be quite different from the actual outcome [11]. This difference represents the effect of actual exposure level relative to exposure level X_j. The only Yij that can be observed is the one corresponding to the exposure actually received by individual i. The remaining Yij cannot be observed but they can be estimated from observed covariates and outcomes [11].

Multiple causal factors can be taken into account in counterfactual models since causal factors may be necessary but not sufficient [16]. By specifying what would happen under alternative possible patterns of exposure, counterfactual models provide useful effect measures for etiologic studies [15]. This allows for the effect measure of interest (the causal contrast) to be more precisely defined than is sometimes the case in epidemiologic studies. A causal contrast compares disease frequency (in a particular target population during a specified time period) under two exposure distributions. As Maldonado and Greenland [15] explained, “A parameter (such as a disease frequency) that describes events under actual conditions is said to be actual (or factual); in contrast, a parameter that describes events under a hypothetical alternative to actual conditions is said to be counterfactual.” Counterfactual parameters cannot be observed since they describe hypothetical events following alternatives to actual conditions [15]. Although critics have argued that the counterfactual model of causation depends on unobservable assumptions [17], the counterfactual approach has been increasingly utilized in epidemiology and other areas of medical and social research.

Greenland and Robins [18] emphasized that the exposure X should be a potentially changeable condition in order to make sense of the unobserved potential outcomes. Although this requirement appears to rule out the application of counterfactual models to most genetic factors, this is not universally true. Counterfactual models could conceivably be applied to epi-genetics data such as patterns of DNA methylation. Other examples of genetics research where counterfactual models of causation have potential are randomized trials involving targeted molecular therapies.

GRAPHICAL CAUSAL MODELS AND STRUCTURAL EQUATIONS MODELING

Several authors have explained how graphical causal models or causal diagrams (including related concepts such as directed acyclic graphs) can be used to better understand causal relationships in health research, epidemiology, and social research [11, 19, 20]. For a recent example, see Hernan and Cole [21]. Although graphical causal models or causal diagrams have mostly been used to identify and explain causal relationships between environmental or social factors and disease outcomes, they also have potential applications in genetics research. For example, in studies of DNA adducts among Army personnel who were likely exposed to smoke from oil well fires during the 1991 Persian Gulf War, graphical causal models could be used to describe the relationships between various environmental sources of polycyclic aromatic hydrocarbons (for example, smoke from oil well fires, cigarette smoking, and dietary sources such as charbroiling of meat) and potential adverse health outcomes such as lung cancer [22, 23]. Potential confounding factors or effect modifiers (for example, age at exposure, sex, time since exposure, and polymorphically expressed genes that code for DNA repair enzymes) might also be represented in the models.

Graphical causal models can be related to structural equation models and to counterfactual models [11, 19]. Structural equations modeling is a multivariate statistical technique in which a web or network of causation is modeled by a system of equations and independence assumptions [11, 24, 25]. Each equation shows how an individual response variable (outcome variable) changes as its direct causal variables change. A variable may appear in no more than one equation as a response variable, but may appear in any other equation as a causal variable [11]. In contrast to ordinary regression equations which represent associations of actual outcomes with actual values of covariates across individuals, structural equations may have within-individual causal interpretations [11]. Several authors have explained the use of structural equations modeling in analyzing health data [11, 19, 24, 25]. Researchers do not always interpret results obtained from structural equations modeling as causal, however [19].

CAUSAL ASSESSMENTS OF RESULTS FROM GENOME-WIDE ASSOCIATION STUDIES

Recent reviews have outlined the rationale and methods for genome-wide association studies [1, 2, 26]. These studies utilize automated, high throughput genotyping technologies to analyze biological specimens collected from cases and controls for up to one million single nucleotide polymorphisms (SNPs) and at relatively low cost. The genome-wide association approach allows interrogation of the entire human genome in large numbers of unrelated individuals [26]. In contrast to studies of candidate genes, there may be no a priori hypotheses about genetic associations with disease in genome-wide association studies. Nearby SNP alleles tend to be inherited together more often than expected by chance because of linkage disequilibrium [1]. Because of the strong associations among the SNPs in most chromosomal regions, only a few SNPs need to be typed to predict the likely variants at the rest of the SNPs in a particular region. As a result, it is not necessary to type all 10 million common SNPs in individuals with and without disease in order to identify sites that differ in frequency between the groups. Studies involving the typing of a few hundred thousand or up to one million of these SNPs can test the hypothesis that one or more common variants explain part of the genetic risk for a disease or trait. If a SNP increases the risk of a common disease, then there will be a statistical association in the population between disease and that SNP and several nearby SNPS due to linkage disequilibrium [2]. Quality control procedures and the replication of results obtained from genome-wide association studies are important considerations, since genotyping errors can cause spurious false positive results or false negative results that obscure true associations [1]. Extensive data cleaning and quality control is required in genome-wide association studies to detect problems that can result in spurious results [2].

Genome-wide association scanning has so far identified variants or chromosomal regions associated with more than 40 complex, non-Mendelian diseases or traits including type 1 and type 2 diabetes, breast, prostate, and colorectal cancer, coronary heart disease and other cardiovascular diseases, neuropsychiatric conditions, and autoimmune and infectious diseases, although the estimated odds ratios are mostly small [1]. In order to have sufficient statistical power to detect genotypic odds ratios on the order of 1.1 to 1.4, it is necessary to include large samples of cases and controls. Replication studies involving other genome-wide association platforms are needed to evaluate the consistency of associations. Studies of disease subtypes or cases with a younger age at onset or more rapidly progressive course are also likely to be informative [1]. The study of rare variants will require large-scale resequencing analyses [2].

Most deleterious variants are found in protein coding regions (exons) or in sequences that control gene expression (promoters). However, gene expression can be altered by SNPs found in noncoding gene segments (introns). Hindorff *et al.* [27] examined characteristics of reported trait and disease-associated SNPs and found that 45% of reported trait/disease-associated SNPs were intronic and 43% were intergenic. Moreover, the trait/disease-associated SNPs were significantly depleted in intergenic regions, which support the assertion that intergenic regions have the smallest ratio of functional to total DNA, even though they may contain important regulatory sequences [27].

Ioannidis *et al.* [14] recently argued that established guidelines for causal inference in epidemiology are inappropriate for assessing genetic associations such as those identified in genome-wide association studies where associations with hundreds of thousands of genetic variants may be examined. While it is true that temporality is not an important consideration for genetic factors fixed at birth, assessments of temporality can be important in studies of epi-genetic influences. Moreover, genetics is not the only subdiscipline within epidemiology where weak associations may be of interest [7]. Because of gene-environment interactions and other factors, the existence and strength of genetic associations may vary across populations and weak associations are frequently identified. In many epidemiologic studies, weak causal effects are often difficult to distinguish from non-causal associations that are due to methodological biases. To overcome such limitations of traditional causal criteria, Ioannidis *et al.* [14] proposed a semi-quantitative index that assigns three levels for the amount of evidence, extent of replication, and protection from bias. Following this approach, a composite assessment of strong, moderate, or weak epidemiologic *credibility* of a genetic association is obtained. In their account, the credibility of a genetic association is improved by consistent evidence obtained from many studies, or a smaller number of large studies. However, the consistency argument is weakened by the fact that gene-environment interactions are known to occur. A potential disadvantage of the use of semi-quantitative indices to assess genetic associations is that they do not assess the concordance of biological and epidemiological evidence. In addition, not all genome-wide association reports provide sufficient detail about epidemiologic study design (for example, complete descriptions of how cases and controls were selected) to evaluate potential biases [26].

Several authors have highlighted the need for replication of results obtained from genome-wide association studies because many initially reported associations were not replicated in subsequent studies [26, 28]. The consistency of observed associations has long been a criterion for causal assessments of results from epidemiologic studies [6]. Genome-wide association studies raise special challenges, however, due to the very large number of tests of disease-SNP associations and the potential for false-positive findings. Biological specimens may be examined to look for patterns of gene expression. The area around a trait/disease-associated SNP may be sequenced to identify rarer variants with more apparent functional significance [26]. Such efforts may be key to understanding the overall biologic coherence of research results.

QUANTITATIVE ANALYSIS OF DATA FROM GENOME-WIDE ASSOCIATION STUDIES

Quantitative methods for analyzing data from genome-wide association studies allow for the identification of statistical associations which may then be replicated in independent samples or validated through studies of gene expression and proteomics. Statistical tests are performed to identify associations between SNPs passing quality thresholds and the disease or trait of interest [29]. Stringent levels of statistical significance are required because of the vast number of tests performed [26]. Biostatistical analyses of data from genome-wide association study data have often been “agnostic” in the sense that they ignore prior knowledge about disease pathobiology [30]. In addition, analyses have often examined associations with one SNP at a time and ignored their genomic context and gene-gene and gene-environment interactions. More recent approaches in bioinformatics strive to take into account non-linear relationships due to gene-gene and gene-environment interactions [30]. Genotypic effects may depend upon environmental exposures. Parametric statistical models which play an important role in contemporary genetic epidemiology have limited power for modeling high-order, non-linear interactions [30]. Efforts are being made to apply data mining techniques, machine learning, and advanced computational methods—for example, random forests, neural networks, and multifactor dimensionality reduction (MDR)—to the immense amounts of data obtained from genome-wide association studies [30, 31]. Such approaches make fewer assumptions about the functional form of the model. Artificial neural networks are discussed below as an example. Researchers are exploring ways to incorporate prior biological knowledge obtained from public databases into data analysis algorithms such as pathway analysis [31, 32]. The success of pathway analysis depends upon the completeness and quality of the information in the public databases.

Although the initial successes of genome-wide association studies have focused attention on genomics, proteomic and transcriptome information are also important. Gene expression data obtained from cell lines and other human tissues (for example, blood, brain, liver) are often key to understanding physiological or regulatory pathways [33, 34]. Studies have shown that quantitative analyses of gene expression data and transcriptome information can help to clarify results obtained from genome-wide association studies [33, 35]. Using data from expression quantitative locus (eQTL) studies, Nicolae *et al.* [35] recently showed that SNPs associated with complex traits and diseases are more likely to be eQTLs than other SNPs chosen from genome-wide association studies (and matched on minor allele frequency). Their results indicate that annotating SNPs with a score reflecting the strength of the evidence that the SNP is an eQTL can enhance the ability to discover true associations in genome-wide association studies and clarify the biological mechanism accounting for the associations [35]. These findings suggest ways to more accurately characterize SNP signals in genome-wide association studies with respect to target genes and biological function. In order to detect differences in gene expression across tissue types, disease, and time, computer algorithms and computational methods have been developed for analyzing data obtained using high-throughput techniques [36, 37]. In order to better understand physiological and pathological processes at the molecular level, pathway analysis and differential network analysis have been applied to detect and quantify signaling cascades and regulatory networks [37].

Artificial Neural Network Analysis

One potential approach for exploring large datasets assembled as part of genome-wide association studies and gene expression and proteomics research include a class of computer intensive techniques known as artificial neural networks [38, 39]. Artificial neural networks are mathematical models that were originally patterned after the structure or functional aspects of biological neural networks [40]. Artificial neural networks are usually adaptive systems that change their structure based upon information that flows through the network during the "learning phase." Functions are performed collaboratively and in parallel by the units or processing elements. In the physical sciences, artificial neural networks have been widely applied to the analysis of complex systems. Artificial neural networks and other "machine learning" techniques have been applied to a wide variety of classification tasks including the analysis of proteomics and gene expression data for cancer patients and cancer cell lines [41-43]. For a review of the advantages and limitations of artificial neural network techniques for the analysis of proteomic data, see Lancashire *et al.* [42].

Neural networks can be contrasted with other statistical techniques such as regression modeling. Such networks automatically allow for nonlinear relations between the independent and dependent variables and all possible interactions between the dependent variables [44]. Artificial neural networks do not require explicit distributional assumptions such as normality [44]. Another advantage of artificial neural network techniques is their ability to handle "noisy" or "fuzzy" information (for example, patterns of diagnostic signatures in proteomic data generated using high throughput mass spectrometry) and they can be used to analyze data with incomplete or missing values [43].

Artificial neural networks do have certain disadvantages. Like other machine learning techniques, artificial neural network approaches require substantial computational resources. In addition, their utility for identifying causal associations in data obtained from genome-wide association studies and proteomics research is still uncertain. They also have a certain "black box" quality [44]. By way of contrast, regression modeling techniques allow for hypothesis testing about the univariate and multivariate associations between each independent variable and the dependent variable of interest [44, 45]. In contrast to neural networks, regression techniques provide information about the relative importance of each explanatory variable [44].

To assess the true predictive performance of a model, an approach to assessing the fit on data that was not used in the model building process is needed. The available approaches include cross-validation, bootstrap validation, and splitting the data set into two samples. Open-source, neural network software is available to research, develop, and apply artificial neural networks and other machine learning techniques [46, 47].

SUMMARY AND CONCLUSIONS

Efforts to identify and quantify causal associations with etiologic agents are a frequent focus of health research which is often multidisciplingary in nature and aimed at more than one disease or adverse health condition. Recent developments in genomics such as the advent of genome-wide association studies have challenged traditional frameworks for causal inference in observational research [14]. However, causal criteria such as the consistency of research findings (i.e., the ability to replicate disease-SNP associations in independent samples) and the overall biological plausibility and biological coherence of research findings are of continuing importance. Traditional epidemiologic approaches for minimizing study bias and for evaluating the potential impact of bias on research findings are important in genome-wide association studies. The temporality of associations identified in epi-genetic studies is also important. Advances in causal analysis such as the counterfactual model and causal diagrams (directed acyclic graphs) may be helpful for conceptualizing the role of epi-genetic factors and gene-environment interactions in studies of disease etiology. In such fields as bioinformatics, the continued development of quantitative techniques for examining gene expression and proteomic data is likely be of considerable value for understanding biologic pathways of disease causation. Quantitative techniques for analyzing data from genome-wide association studies, combined with gene expression and environmental exposure data, may be enhanced by mathematical and statistical tools for exploring causal questions and identifying causal relationships in health data. In addition to replication of findings in independent samples, results obtained from genome-wide association studies may be bolstered by experimental studies that examine the functional implications of disease-SNP associations.

REFERENCES

[1] Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. J Clin Investig 2008; 118: 1590-605.

[2] Psychiatric GWAS Consortium Coordinating Committee. Genomewide association studies: history, rationale, and prospects for psychiatric disorders. Am J Psychiatry 2009; 166: 540-56.

[3] Ward A. Causal criteria and the problem of complex causation. Med Health Care Philos 2009; 12: 333-43.

[4] Rothman KJ, Greenland S. Causation and causal inference in epidemiology. Am J Public Health 2005; 95: S144-S50.

[5] Hill AB. The environment and disease: association or causation? Proc R Soc Med 1965; 58: 295-300.

[6] Susser M. What is a cause and how do we know one? A grammar for pragmatic epidemiology. Am J Epidemiol 1991; 133: 635-48.

[7] Coughlin SS. Causal inference and scientific paradigms in epidemiologic research. Bentham e-Books 2010. http://bentham.org/ebooks/9781608051816/index.htm

[8] Rothman KM. Causal inference in epidemiology. In: Modern epidemiology. Boston: Little, Brown and Company 1986.

[9] Greenland S. An overview of methods for causal inference from observational studies. In: Gelman A, Meng XL, Eds. Applied bayesian modeling and causal inference from an incomplete-data perspective. New York: Wiley 2004.

[10] Greenland S. Causal analysis in the health sciences. J Am Stat Assoc 2000; 95: 286-9.

[11] Greenland S, Brumback B. An overview of relations among causal modeling methods. Int J Epidemiol 2002; 31: 1030-7.

[12] Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies *via* potential outcomes: concepts and analytical approaches. Annu Rev Public Health 2000; 21: 121-45.

[13] Robins J, Greenland S. The probability of causation under a stochastic model for individual risk. Biometrics 1989; 45: 1125-38.

[14] Ioannidis JP, Boffetta P, Little J, *et al.* Assessment of cumulative evidence on genetic associations: interim guidelines. Int J Epidemiol 2008; 37: 120-32.

[15] Maldonado G, Greenland S. Estimating causal effects. Int J Epidemiol 2002; 31: 422-9.

[16] Hofler M. Causal inference based on counterfactuals. BMC Med Res Methodol 2005; 5: 28.

[17] Dawid AP. Causal thinking without counterfactuals. J Am Stat Assoc 2000; 95: 407-24.

[18] Greenland S, Robins JM. Identifiability, exchangeability and confounding revisited. Epidemiol Perspect Innov 2009; 6: 4.

[19] Pearl J. Causality, models, reasoning and inference. New York: Springer 2000.

[20] Morgan SL, Winship C. Counterfactuals and causal inference. Methods and principles for social research. New York: Cambridge University Press 2007

[21] Hernan MA, Cole SR. Causal diagrams and measurement bias. Am J Epidemiol 2009; 170: 959-62.

[22] Poirier MC, Weston A, Schoket B, *et al.* Biomonitoring of United States army soldiers serving in Kuwait in 1991. Cancer Epidemiol Biomarkers Prev 1998; 7: 545-51.

[23] Schoket B. DNA damage in humans exposed to environmental and dietary polycyclic aromatic hydrocarbons. Mutat Res 1999; 424: 143-53.

[24] Goldberger AS. Structural equation models in the social sciences. Econometrica 1972; 40: 979-1001.

[25] Duncan OD. Introduction to structural equation models. New York: Academic Press 1975.

[26] Pearson TA, Manolio TA. How to interpret a genome-wide association study. JAMA 2008; 299: 1335-44.

[27] Hindorff LA, Sethupathy P, Junkins HA, *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 2009; 106: 9362-7.

[28] Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. Genet Med 2002; 4: 45-61.

[29] Ziegler A, Konig IR, Thompson JR. Biostatistical aspects of genome-wide association studies. Biomed J 2008; 50: 8-28.

[30] Moore JH, Williams SM. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. Bioessays 2005; 27: 637-46.

[31] Sebastiani P, Timofeev N, Dworkis DA, *et al.* Genome-wide association studies and the genetic dissection of complex traits. Am J Hematol 2009; 84: 504-15.

[32] Elbers CC, van Eijk KR, Franke L, *et al.* Using genome-wide pathyway analysis to unravel the etiology of complex diseases. Genet Epidemiol 2009; 33: 419-31.

[33] Schadt EE, Lamb J, Yang X, *et al.* An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet 2005; 37: 710-7.

[34] Dermitzakis ET. From gene expression to disease risk. Nat Genet 2008; 40: 492-3.

[35] Nicolae DL, Gamazon E, Zhang W, *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet 2010: 6: e1000888.

[36] Degenhordt J, Haubrock M, Donitz J, *et al.* DEEP-a tool for differential expression effectors prediction. Nucleic Acids Res 2007; 35: W619-24.

[37] Keller A, Backes C, Gerasch A, *et al.* A novel algorithm for detecting differentially regulated paths based on gene set enrichment analysis. Bioinformatics 2009; 25: 2787-94.

[38] Dreiselti S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform 2002; 35: 352-9.

[39] Ripley BD. Pattern recognition and neural networks. Cambridge, UK: Cambridge University Press 1996.

[40] Wikipedia. Available from: http://en.wikipedia.org/wiki/ArtificialVneuralVnetwork [cited: 21st Dec 2009].

[41] Johann DJ, Jr., McGuigan MD, Tomov S, *et al.* Novel approaches to visualization and data mining reveals diagnostic information in the low amplitude region of serum mass spectra from ovarian cancer patients. Dis Markers 2004; 19: 197-207.

[42] Lancashire LJ, Rees RC, Ball GR. Identification of gene transcript signatures predictive for estrogen receptor and lymph node status using a stepwise forward selection artificial neural network modeling approach. Artif Intell Med 2008; 43: 99-111.

[43] Lancashire LJ, Mian S, Ellis IO, *et al.* Current developments in the analysis of proteomic data: artificial neural network data mining techniques for the identification of proteomic biomarkers related to breast cancer. Curr Proteomics 2005; 2: 15-29.

[44] Sargent DJ. Comparison of artificial neural networks with other statistical approaches. Results from medical data sets. Cancer 2001; 91: 1636-42.

[45] Warner B, Misra M. Understanding neural networks as statistical tools. Am Stat 1996; 50: 284-93.

[46] Zupan B, Demsar J. Open-source tools for data mining. Clin Lab Med 2008; 28: 37-54.

[47] Frank E, Hall M, Homes G, Witten IH. Data mining in bioinformatics using Weka. Bioinformatics 2004; 20: 2479-81.

Steven S. Coughlin
(***Editor-in-Chief***)
Environmental Epidemiology Service (135)
Office of Public Health and Environmental Hazards
Department of Veterans Affairs
810 Vermont Ave., NW
Washington, DC 20420
USA
Tel: (202) 266-4656
Fax: (202) 495-5956
E-mail: steven.coughlin@va.gov