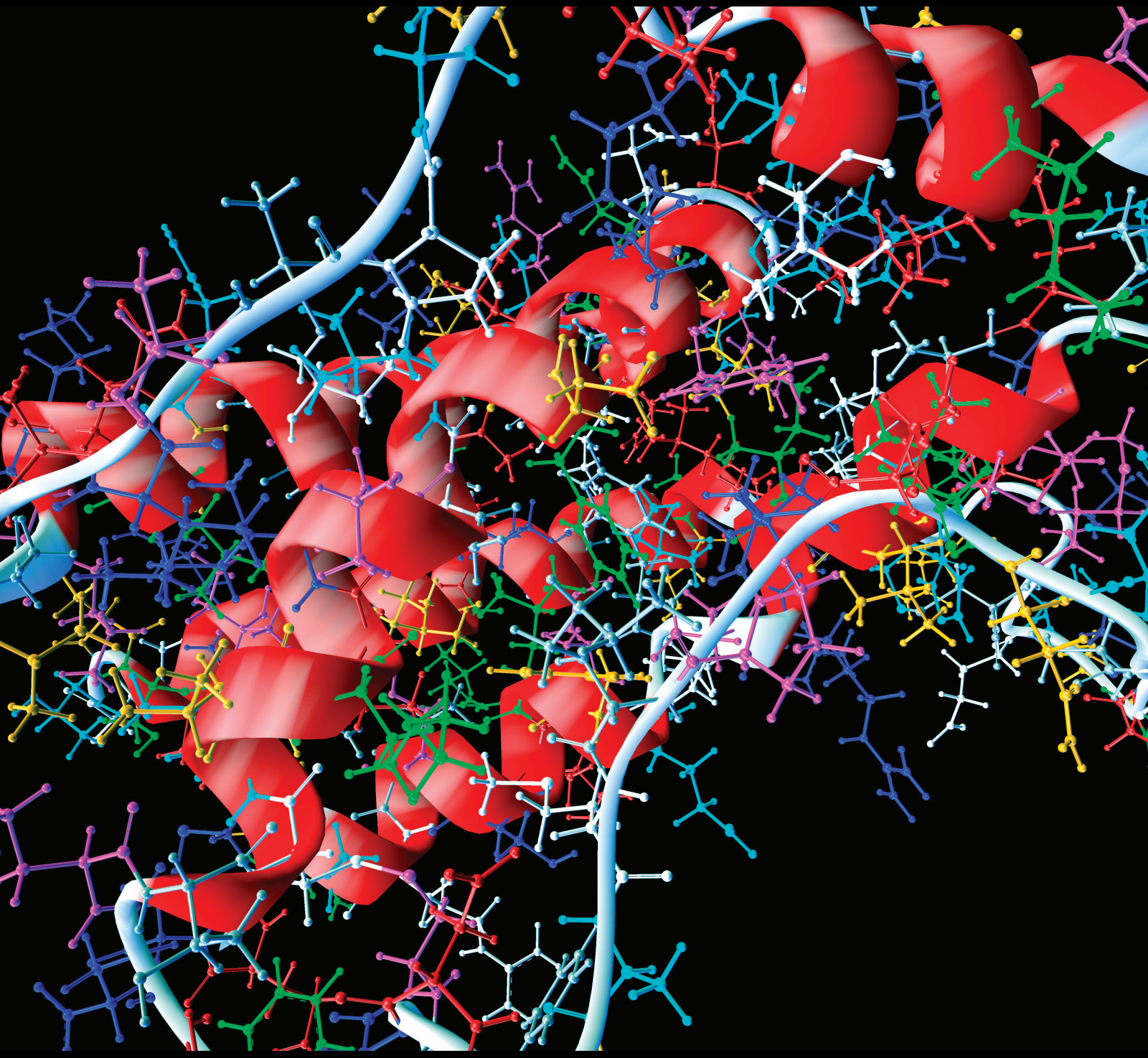
Computational and Mathematical Methods in Medicine

Advances in Statistical Medicine

Guest Editors: Sujay Datta, Xiao-Qin Xia, Samsiddhi Bhattacharjee, and Zhenyu Jia



Advances in Statistical Medicine

Computational and Mathematical Methods in Medicine

Advances in Statistical Medicine

Guest Editors: Sujay Datta, Xiao-Qin Xia,
Samsiddhi Bhattacharjee, and Zhenyu Jia

Editorial Board

Contents

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 316153, 2 pages
http://dx.doi.org/10.1155/2014/316153

Editorial
Advances in Statistical Medicine

Sujay Datta,[1] Xiao-Qin Xia,[2] Samsiddhi Bhattacharjee,[3] and Zhenyu Jia[1,4]

[1] *Department of Statistics, University of Akron, 302 Buchtel Commons, Akron, OH 44325, USA*
[2] *Institute of Hydrobiology, Chinese Academy of Sciences, No. 7 Donghu South Road, Wuhan, Hubei 430072, China*
[3] *National Institute of Biomedical Genomics, Netaji Subhas Sanatorium, 2nd Floor, P.O. Box N.S.S., Kalyani, West Bengal 741251, India*
[4] *Department of Family and Community Medicine, Northeast Ohio Medical University, 4209 Ohio 44, Rootstown, OH 44272, USA*

Correspondence should be addressed to Zhenyu Jia; zjia@uakron.edu

Received 17 September 2014; Accepted 17 September 2014; Published 9 November 2014

We are not even halfway through the second decade of the 21st century and there is already ample evidence that it is going to be the century of biotechnology, leading to unprecedented breakthroughs in the medical sciences and revolutionizing everything from drug discovery to healthcare delivery. The rapid advancement in high-performance computing that took place in the last quarter of the last century has been a key driving force in this revolution, enabling us to generate, store, query, and transfer huge amounts of medical data. This is where statisticians come into the picture, lending their expertise in extracting information from data and converting that information to medical knowledge.

The crucial role that statisticians have been playing in this information revolution has created new challenges and posed difficult problems for their own discipline. Dealing with them has often necessitated new statistical techniques, new approaches to inference, or even new modes of thinking. These, in turn, have been the motivating force behind an astonishing flurry of biostatistical research activities in the recent years. In the ten carefully chosen and peer-reviewed articles of this special issue, we hope to provide a nuanced perspective on some of the areas in the biomedical sciences that have directly benefited from that research. This thriving partnership between experts in the quantitative world and those in the medical world has been highly interdisciplinary in nature. This special issue aims to introduce researchers, practitioners, and students on both sides of the fence to some of the statistical modeling and inference approaches that have collectively had such a huge impact on the field of medicine. And there is a clear need for it.

Due to the injection of a steady flow of new technologies, the medical field has progressed rapidly and has produced data at a phenomenal rate. It is important for those in the medical world to understand that the types of data collected and the manner in which they are collected are crucial to the validity and reliability of the subsequent statistical analysis. Some basic familiarity with statistical methodologies will make them aware of the potential pitfalls of some designs of experiments in certain contexts and enable them to choose better ones. Also, they need to realize that statistical analysis is not a mechanical process like solving a set of mathematical equations. Specifying a statistical model that is appropriate for a given situation and drawing conclusions about the model parameters are fraught with many challenges. This realization will give them a better appreciation of the role that a statistician plays in medical research. On the other hand, statisticians will be motivated to develop methodologies capable of handling systems that change constantly with time and in response to therapeutic, physiological, and environmental stimuli. They will see the need for dealing with mathematical models that are much more complex and challenging than those routinely encountered in the rest of statistics.

The articles in this special issue were chosen with this in mind. J. V. Pottala et al. use a latent variable approach and structural equation modeling for analyzing erythrocyte fatty acids in the context of the Framingham study. B. Miller et al. use chi-squared automatic interaction detection decision trees and waist circumference as a surrogate measure to detect metabolic syndrome in young adults. M. Banerjee et al. use

logic regression in an innovative way in the context of kidney cancer treatment delivery to uncover the complex interplay among patient, provider, and practice environment variables based on linked data from the National Cancer Institute's Surveillance, Epidemiology and End Results Program and Medicare. Z. Huang and Y. Chen propose and implement a two-stage model based on synergetic neural networks for exon recognition, a fundamentally important task in bioinformatics. J. A. Koziol and Z. Jia generalize the quadratic version of the log-rank test, introduced originally by Lin and Wang, to incorporate weights that increase statistical power in some situations. H. Li et al. construct an association network between micro-RNA and cancer based on more than a thousand miRNA-cancer associations detected from millions of abstracts using a text-mining method. C. Taslim and S. Lin propose a mixture modeling framework that is flexible enough to automatically adapt to most high-throughput data-types that are encountered in modern genomics, thereby overcoming the difficulty that statistical methods specifically designed for one data-type may not be optimal for or applicable to another data-type. P. T. Edlefsen shows through examples that the heterogeneous effects of leaky vaccines (that protect subjects with fewer exposures to a pathogen at a higher effective rate than subjects with more exposures) violate the proportional hazards assumption, leading to incomparability of infected cases across treatment groups and to nonindependence of the distributions of the competing failure processes in a competing risks setting. E. Cheng and Z. M. Ozsoyoglu propose a framework for deriving path-counting formulas for all generalized kinship coefficients for which there are recursive formulas and which are sufficient for computing condensed identity coefficients, an important computation on Pedigree data that provides a complete description of the degree of relatedness between two individuals. Finally, R. L. Einsporn and Z. Jia shed the light on some problems with interaction and varying block sizes in a comparison of endotracheal tubes through a randomized clinical experiment based on a block design.

Collectively, the editors express their sincerest gratitude to their respective institutions for the time and resources that they were provided. And last but not least, the editors gratefully acknowledge the support and encouragement they received from their respective families during this endeavor.

Sujay Datta
Xiao-Qin Xia
Samsiddhi Bhattacharjee
Zhenyu Jia

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 898424, 20 pages
http://dx.doi.org/10.1155/2014/898424

Research Article

Path-Counting Formulas for Generalized Kinship Coefficients and Condensed Identity Coefficients

En Cheng[1] and Z. Meral Ozsoyoglu[2]

[1] *Computer Science Department, The University of Akron, Akron, OH 44325, USA*
[2] *Electrical Engineering and Computer Science Department, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106, USA*

Correspondence should be addressed to En Cheng; echeng@uakron.edu

Received 14 January 2014; Accepted 8 May 2014; Published 21 July 2014

Academic Editor: Zhenyu Jia

An important computation on pedigree data is the calculation of condensed identity coefficients, which provide a complete description of the degree of relatedness of two individuals. The applications of condensed identity coefficients range from genetic counseling to disease tracking. Condensed identity coefficients can be computed using linear combinations of generalized kinship coefficients for two, three, four individuals, and two pairs of individuals and there are recursive formulas for computing those generalized kinship coefficients (Karigl, 1981). Path-counting formulas have been proposed for the (generalized) kinship coefficients for two (three) individuals but there have been no path-counting formulas for the other generalized kinship coefficients. It has also been shown that the computation of the (generalized) kinship coefficients for two (three) individuals using path-counting formulas is efficient for large pedigrees, together with path encoding schemes tailored for pedigree graphs. In this paper, we propose a framework for deriving path-counting formulas for generalized kinship coefficients. Then, we present the path-counting formulas for all generalized kinship coefficients for which there are recursive formulas and which are sufficient for computing condensed identity coefficients. We also perform experiments to compare the efficiency of our method with the recursive method for computing condensed identity coefficients on large pedigrees.

1. Introduction

With the rapidly expanding field of medical genetics and genetic counseling, genealogy information is becoming increasingly abundant. In January 2009, the US Department of Health and Human Services released an updated and improved version of the Surgeon General's Web-based family health history tool [1]. This Web-based tool makes it easy for users to record their family health history. Large extended human pedigrees are very informative for linkage analysis. Pedigrees including thousands of members in 10–20 generations are available from genetically isolated populations [2, 3]. In human genetics, a pedigree is defined as "a simplified diagram of a family's genealogy that shows family members' relationships to each other and how a specific trait, abnormality, or disease has been inherited" [4]. Pedigrees are utilized to trace the inheritance of a specific disease, calculate genetic risk ratios, identify individuals at risk, and facilitate genetic counseling. To calculate genetic risk ratios or identify individuals at risk, we need to assess the degree of relatedness of two individuals. As a matter of fact, all measures of relatedness are based on the concept of *identical by descent* (IBD). Two alleles are identical by descent if one is an ancestral copy of the other or if they are both copies of the same ancestral allele. The IBD concept is primarily due to Cotterman [5] and Malecot [6] and has been successfully applied to many problems in population genetics.

The simplest measure of relationship between two individuals is their kinship coefficient. The *kinship coefficient* between two individuals i and j is the probability that an allele selected randomly from i and an allele selected randomly from the same autosomal locus of j are identical by descent. To better discriminate between different types of pairs of relatives, identity coefficients were introduced by Gillois [7] and

Harris [8] and promulgated by Jacquard [9]. Considering the four alleles of two individuals at a fixed autosomal locus, there are 15 possible identity states. Disregarding the distinction between maternally and paternally derived alleles, we obtain 9 condensed identity states. The probabilities associated with each condensed identity state are called *condensed identity coefficients*, which are useful in a diverse range of fields. This includes the calculation of risk ratios for qualitative disease, the analysis of quantitative traits, and genetic counseling in medicine.

A recursive algorithm for calculating condensed identity coefficients proposed by Karigl [10] has been known for some time. This method requires that one calculates a set of generalized kinship coefficients, from which one obtains condensed identity coefficients via a linear transformation. One limitation is that this recursive approach is not scalable when applied to very large pedigrees. It has been previously shown that the kinship coefficients for two individuals [11–13] and the generalized kinship coefficients for three individuals [14, 15] can be efficiently calculated using path-counting formulas together with path encoding schemes tailored for pedigree graphs.

Motivated by the efficiency of path-counting formulas for computing the kinship coefficient for two individuals and the generalized kinship coefficient for three individuals, we first introduce a framework for developing path-counting formulas to compute generalized kinship coefficients concerning three individuals, four individuals, and two pairs of individuals. Then, we present path-counting formulas for all generalized kinship coefficients which have recursive formulas proposed by Karigl [10] and are sufficient to compute condensed identity coefficients. In summary, our ultimate goal is to use path-counting formulas for generalized kinship coefficients computation so that efficiency and scalability for condensed identity coefficients calculation can be improved.

The main contributions of our work are as follows:

(i) a framework to develop path-counting formulas for generalized kinship coefficients;

(ii) a set of path-counting formulas for all generalized kinship coefficients having recursive formulas [10];

(iii) experimental results demonstrating significant performance gains for calculating condensed identity coefficients based on our proposed path-counting formulas as compared to using recursive formulas [10].

2. Materials and Methods

This section describes kinship coefficients and generalized kinship coefficients, identity coefficients, and condensed identity coefficients in more detail. Conceptual terms for the path-counting formulas for three and four individuals are introduced in Section 2.3. In addition, an overview of path-counting formula derivation is presented.

2.1. Kinship Coefficients and Generalized Kinship Coefficients. The kinship coefficient between two individuals a and b is the probability that a randomly chosen allele at the same locus from each is identical by descent (IBD). There are two approaches to computing the kinship coefficient Φ_{ab}: the recursive approach [10] and the path-counting approach [16]. The recursive formulas [10] for Φ_{ab} and Φ_{aa} are

$$\Phi_{ab} = \frac{1}{2}\left(\Phi_{fb} + \Phi_{mb}\right) \quad \text{if } a \text{ is not an ancestor of } b,$$
$$\Phi_{aa} = \frac{1}{2}\left(1 + \Phi_{fm}\right) = \frac{1}{2}\left(1 + F_a\right), \tag{1}$$

where f and m denote the father and the mother of a, respectively, and F_a is the inbreeding coefficient of a.

Wright's path-counting formula [16] for Φ_{ab} is

$$\Phi_{ab} = \sum_{A} \sum_{\langle P_{Aa}, P_{Ab}\rangle \in PP} \left(\frac{1}{2}\right)^{r+s+1} \left(1 + F_A\right), \tag{2}$$

where A is a common ancestor of a and b, PP is a set of nonoverlapping path-pairs $\langle P_{Aa}, P_{Ab}\rangle$ from A to a and b, r is the length of the path P_{Aa}, s is the length of the path P_{Ab}, and F_A is the inbreeding coefficient of A. The path-pair $\langle P_{Aa}, P_{Ab}\rangle$ is *nonoverlapping* if and only if the two paths share no common individuals, except A.

Recursive formulas proposed by Karigl [10] for generalized kinship coefficients concerning three individuals, four individuals, and two pairs of individuals are listed as follows in (3), (4), and (5):

$$\Phi_{abc} = \frac{1}{2}\left(\Phi_{fbc} + \Phi_{mbc}\right)$$
$$\text{if } a \text{ is not an ancestor of } b \text{ or } c,$$
$$\Phi_{aab} = \frac{1}{2}\left(\Phi_{ab} + \Phi_{fmb}\right) \quad \text{if } a \text{ is not an ancestor of } b,$$
$$\Phi_{aaa} = \frac{1}{4}\left(1 + 3\Phi_{fm}\right) = \frac{1}{4}\left(1 + 3F_a\right), \tag{3}$$

$$\Phi_{abcd} = \frac{1}{2}\left(\Phi_{fbcd} + \Phi_{mbcd}\right)$$
$$\text{if } a \text{ is not an ancestor of } b \text{ or } c \text{ or } d,$$
$$\Phi_{aabc} = \frac{1}{2}\left(\Phi_{abc} + \Phi_{fmbc}\right)$$
$$\text{if } a \text{ is not an ancestor of } b \text{ or } c, \tag{4}$$
$$\Phi_{aaab} = \frac{1}{4}\left(\Phi_{ab} + 3\Phi_{fmb}\right)$$
$$\text{if } a \text{ is not an ancestor of } b,$$
$$\Phi_{aaaa} = \frac{1}{8}\left(1 + 7\Phi_{fm}\right) = \frac{1}{8}\left(1 + 7F_a\right),$$

$$\Phi_{ab,cd} = \frac{1}{2}\left(\Phi_{fb,cd} + \Phi_{mb,cd}\right)$$

if a is not an ancestor of b or c or d,

$$\Phi_{aa,bc} = \frac{1}{2}\left(\Phi_{bc} + \Phi_{fm,bc}\right)$$

if a is not an ancestor of b or c,

$$\Phi_{ab,ac} = \frac{1}{4}\left(2\Phi_{abc} + \Phi_{fb,mc} + \Phi_{mb,fc}\right)$$

if a is not an ancestor of b or c,

$$\Phi_{aa,ab} = \frac{1}{4}\left(\Phi_{ab} + \Phi_{fmb}\right)$$

if a is not an ancestor of b,

$$\Phi_{aa,aa} = \frac{1}{4}\left(1 + 3\Phi_{fm}\right) = \frac{1}{4}\left(1 + 3F_a\right). \tag{5}$$

Φ_{abc} is the probability that randomly chosen alleles at the same locus from each of the three individuals (i.e., a, b, and c) are identical by descent (IBD). Similarly, Φ_{abcd} is the probability that randomly chosen alleles at the same locus from each of the four individuals (i.e., a, b, c, and d) are IBD. $\Phi_{ab,cd}$ is the probability that a random allele from a is IBD with a random allele from b and that a random allele from c is IBD with a random allele from d at the same locus. Note that $\Phi_{abc} = 0$ if there is no common ancestor of a, b, and c. $\Phi_{abcd} = 0$ if there is no common ancestor of a, b, c, and d, and $\Phi_{ab,cd} = 0$ in the absence of a common ancestor either for a and b or for c and d.

2.2. Identity Coefficients and Condensed Identity Coefficients. Given two individuals a and b with maternally and paternally derived alleles at a fixed autosomal locus, there are 15 possible identity states, and the probabilities associated with each identity state are called *identity coefficients*. Ignoring the distinction between maternally and paternally derived alleles, we categorize the 15 possible states to 9 condensed identity states, as shown in Figure 1. The states range from state 1, in which all four alleles are IBD, to state 9, in which none of the four alleles are IBD. The probabilities associated with each condensed identity state are called *condensed identity coefficients*, denoted by $\{\Delta_i \mid 1 \le i \le 9\}$. The condensed identity coefficients can be computed based on generalized kinship coefficients using the linear transformation shown as follows in (6):

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 & 1 & 1 & 1 & 1 & 1 \\ 2 & 2 & 1 & 1 & 2 & 2 & 1 & 1 & 1 \\ 4 & 0 & 2 & 0 & 2 & 0 & 2 & 1 & 0 \\ 8 & 0 & 4 & 0 & 2 & 0 & 2 & 1 & 0 \\ 8 & 0 & 2 & 0 & 4 & 0 & 2 & 1 & 0 \\ 16 & 0 & 4 & 0 & 4 & 0 & 2 & 1 & 0 \\ 4 & 4 & 2 & 2 & 2 & 2 & 1 & 1 & 1 \\ 16 & 0 & 4 & 0 & 4 & 0 & 4 & 1 & 0 \end{bmatrix} \begin{bmatrix} \Delta_1 \\ \Delta_2 \\ \Delta_3 \\ \Delta_4 \\ \Delta_5 \\ \Delta_6 \\ \Delta_7 \\ \Delta_8 \\ \Delta_9 \end{bmatrix} = \begin{bmatrix} 1 \\ 2\Phi_{aa} \\ 2\Phi_{bb} \\ 4\Phi_{ab} \\ 8\Phi_{aab} \\ 8\Phi_{abb} \\ 16\Phi_{aabb} \\ 4\Phi_{aa,bb} \\ 16\Phi_{ab,ab} \end{bmatrix}. \tag{6}$$

In our work, we focus on deriving the path-counting formulas for the generalized kinship coefficients, including Φ_{abc}, Φ_{abcd}, and $\Phi_{ab,cd}$.

2.3. Terms Defined for Path-Counting Formulas for Three and Four Individuals

(1) Triple-Common Ancestor. Given three individuals a, b, and c, if A is a common ancestor of the three individuals, then we call A a *triple-common ancestor* of a, b, and c.

(2) Quad-Common Ancestor. Given four individuals a, b, c, and d, if A is a common ancestor of the four individuals, then we call A a *quad-common ancestor* of a, b, c, and d.

(3) $P(A, a)$. It denotes the set of all possible paths from A to a, where the paths can only traverse edges in the direction of parent to child such that $P(A, a) \neq NULL$ if and only if A is an ancestor of a. P_{Aa} denotes a particular path from A to a, where $P_{Aa} \in P(A, a)$.

(4) Path-Pair. It consists of two paths, denoted as $\langle P_{Aa}, P_{Ab}\rangle$, where $P_{Aa} \in P(A, a)$ and $P_{Ab} \in P(A, b)$.

(5) Nonoverlapping Path-Pair. Given a path-pair $\langle P_{Aa}, P_{Ab}\rangle$, it is *nonoverlapping* if and only if the two paths share no common individuals, except A.

(6) Path-Triple. It consists of three paths, denoted as $\langle P_{Aa}, P_{Ab}, P_{Ac}\rangle$, where $P_{Aa} \in P(A, a)$, $P_{Ab} \in P(A, b)$, and $P_{Ac} \in P(A, c)$.

(7) Path-Quad. It consists of four paths, denoted as $\langle P_{Aa}, P_{Ab}, P_{Ac}, P_{Ad}\rangle$, where $P_{Aa} \in P(A, a)$, $P_{Ab} \in P(A, b)$, $P_{Ac} \in P(A, c)$, and $P_{Ad} \in P(A, d)$.

(8) $Bi\_C(P_{Aa}, P_{Ab})$. It denotes all common individuals shared between P_{Aa} and P_{Ab}, except A.

(9) $Tri\_C(P_{Aa}, P_{Ab}, P_{Ac})$. It denotes all common individuals shared among P_{Aa}, P_{Ab}, and P_{Ac}, except A.

(10) $Quad\_C(P_{Aa}, P_{Ab}, P_{Ac}, P_{Ad})$. It denotes all common individuals shared among P_{Aa}, P_{Ab}, P_{Ac}, and P_{Ad}, except A.

(11) Crossover and 2-Overlap Individual. If $s \in Bi\_C(P_{Aa}, P_{Ab})$, we call s a *crossover* individual with respect to P_{Aa} and P_{Ab} if the two paths pass through *different* parents of s. On the other hand, if P_{Aa} and P_{Ab} pass through the *same* parent of s, then we call s a 2-*overlap* individual with respect to P_{Aa} and P_{Ab}.

(12) 3-Overlap Individual. If $s \in Tri\_C(P_{Aa}, P_{Ab}, P_{Ac})$ and the three paths P_{Aa}, P_{Ab}, and P_{Ac} pass through the *same* parent of s, then we call s a *3-overlap individual* with respect to P_{Aa}, P_{Ab}, and P_{Ac}.

(13) 2-Overlap Path. If s is a 2-overlap individual with respect to P_{Aa} and P_{Ab}, then both P_{Aa} and P_{Ab} pass through the same parent of s, denoted by p, and the edge from p to s is called an *overlap* edge. All consecutive overlap edges constitute a path and this path is called a 2-*overlap path*. If the 2-overlap path

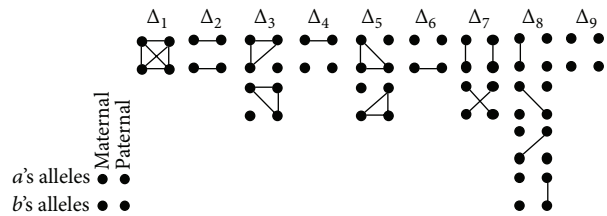
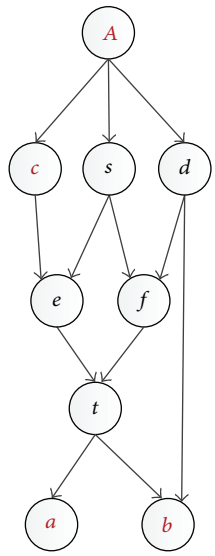


Figure 1: The 15 possible identity states for individuals a and b, grouped by their 9 condensed states. Lines indicate alleles that are IBD.



Path-pair1 $\begin{cases} A \to s \to e \to t \to a \\ A \to s \to e \to t \to b \end{cases}$

where $\{s, e, t\}$ are 2-overlap individuals.

Path-pair2 $\begin{cases} A \to s \to e \to t \to a \\ A \to d \to b \end{cases}$

Non-overlapping path-pair

Path-triple1 $\begin{cases} A \to s \to e \to t \to a \\ A \to d \to f \\ A \to c \end{cases}$

Three independent paths

Path-triple2 $\begin{cases} A \to s \to e \to t \to a \\ A \to d \to f \to t \to b \\ A \to c \end{cases}$

where t is a crossover individual

Path-triple3 $\begin{cases} A \to s \to e \to t \to a \\ A \to s \to e \to t \to b \\ A \to c \end{cases}$

where $\{s, e, t\}$ are 2-overlap individuals and the overlap path is a root 2-overlap path.

Path-pair3 $\begin{cases} A \to s \to e \to t \to a \\ A \to d \to f \to t \to b \end{cases}$

where t is a crossover individual.

Path-pair4 $\begin{cases} A \to c \to t \to e \to a \\ A \to d \to f \to t \to b \end{cases}$

where t is a 2-overlap individual and e is a crossover individual.

Path-triple4 $\begin{cases} A \to s \to e \to t \to a \\ A \to s \to f \to t \to b \\ A \to c \end{cases}$

where t is a crossover individual; s is a 2-overlap individual and the overlap path is a root 2-overlap path.

Path-triple5 $\begin{cases} A \to c \to e \to t \to a \\ A \to c \to e \to t \to b \\ A \to c \end{cases}$

where c is a 3-overlap individual; and $\{e, t\}$ are 2-overlap individuals and the overlap path is a root 2-overlap path

Path-triple6 $\begin{cases} A \to c \to e \to t \to a \\ A \to c \to e \to t \to b \\ A \to c \end{cases}$

where e is a crossover individual; t is a 2-overlap individual and the overlap path is not a root 2 -overlap path; c is a 2-overlap individual and the overlap path is a root 2-overlap path

Figure 2: Examples of path-pairs and path-triples.

extends all the way to the ancestor A, we call it a *root 2-overlap path*.

(14) 3-Overlap Path. It consists of all 3-overlap individuals in a consecutive order. If the 3-overlap path extends all the way to the root A, we call it a *root 3-overlap path*.

Example 1. Consider the *path-pairs* from A to a and b in Figure 2, where A is a common ancestor of a and b. For *path-pair*1, $Bi\_C(P_{Aa}, P_{Ab}) = \{s, e, t\}$, and $A \to s \to e \to t$ is a *root 2-overlap* path with respect to P_{Aa} and P_{Ab}. For *path-pair*4, $Bi\_C(P_{Aa}, P_{Ab}) = \{e, t\}$, where e is a *crossover* individual; t is a 2-*overlap individual* with respect to P_{Aa} and P_{Ab}, and $e \to t$ is a *root 2-overlap* path with respect to P_{Aa} and P_{Ab}.

Example 2. There are four path-quads listed in Figure 3, from A to four individuals a, b, c, and d, where A is a quad-common ancestor of the four individuals. For *path-quad2*, considering the paths P_{Aa} and P_{Ab}, the path $A \to t \to f \to s$ is a *root* 2-*overlap path*; $\{t, f, s\}$ are 2-*overlap* individuals with respect to P_{Aa} and P_{Ab}. For *path-quad3*, $\{t, f, s\}$ are 3-*overlap* individuals with respect to P_{Aa}, P_{Ab}, and P_{Ac}, and the path $A \to t \to f \to s$ is a *root 3-overlap path*.

Then, we summarize all the conceptual terms used in the path-counting formulas for two individuals, three individuals, and four individuals in Table 1 which reveals a glimpse of our framework for generalizing Wright's formula to three and four individuals from terminology aspect.

2.4. An Overview of Path-Counting Formula Derivation. According to Wright's path-counting formula [16] (see (2)) for two individuals a and b, the path-counting approach requires identifying common ancestors of a and b and calculating the contribution of each common ancestor to Φ_{ab}. More specifically, for each common ancestor, denoted as A, we obtain all path-pairs from A to a and b and identify acceptable path-pairs. For Φ_{ab}, an acceptable path-pair $\langle P_{Aa}, P_{Ab} \rangle$ is a nonoverlapping path-pair where

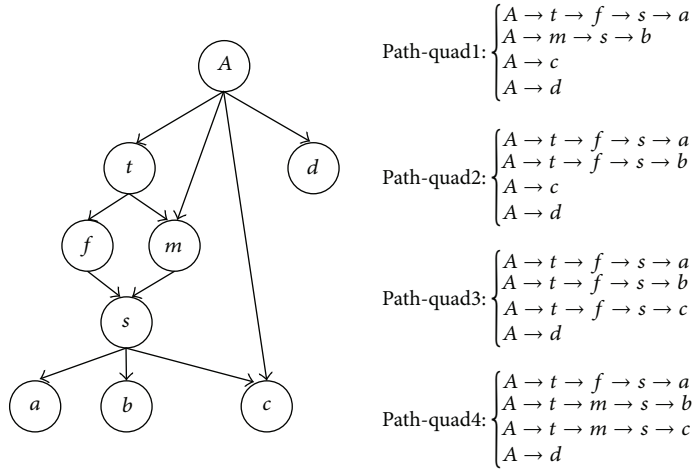


Figure 3: Examples of path-quads.

Table 1: The conceptual terms used for two, three, and four individuals.

Two individuals	Three individuals	Four individuals
Common ancestor	Triple-common ancestor	Quad-common ancestor
Path-pair	Path-triple	Path-quad
$Bi\_C(P_{Aa}, P_{Ab})$	$Tri\_C(P_{Aa}, P_{Ab}, P_{Ac})$	$Quad\_C(P_{Aa}, P_{Ab}, P_{Ac}, P_{Ad})$
N/A	2-Overlap individual	3-Overlap individual
N/A	2-Overlap path	3-Overlap path
N/A	Root 2-overlap path	Root 3-overlap path
N/A	Crossover individual	Crossover individual

the two paths share no common individuals, except A. In Figure 2, *path-pair2* is an acceptable path-pair, while *path-pair1*, *path-pair3*, and *path-pair4* are not acceptable path-pairs. The contribution of each common ancestor A to Φ_{ab} is computed based on the inbreeding coefficient of A, modified by the length of each acceptable path-pair.

To compute Φ_{abc}, the path-counting approach requires identifying all triple-common ancestors of a, b, and c and summing up all triple-common ancestors' contributions to Φ_{abc}. For each triple-common ancestor, denoted as A, we first identify all path-triples each of which consists of three paths from A to a, b, and c, respectively. Some examples of path-triples are presented in Figure 2.

For Φ_{ab}, only nonoverlapping path-pairs are acceptable. A path-triple $\langle P_{Aa}, P_{Ab}, P_{Ac}\rangle$ consists of three path-pairs $\langle P_{Aa}, P_{Ab}\rangle$, $\langle P_{Aa}, P_{Ac}\rangle$, and $\langle P_{Ab}, P_{Ac}\rangle$. For Φ_{abc}, a path-triple might be acceptable even though either 2-overlap individuals or crossover individuals exist between a path-pair. The main challenge we need to address is finding necessary and sufficient conditions for acceptable path-triples.

Aiming at solving the problem of identifying acceptable path-triples, we first use a systematic method to generate all possible cases for a path-pair by considering different types of common individuals shared between the two paths. Then, we introduce building blocks which are connected graphs with conditions on every edge in the graph that encapsulates a set of acceptable cases of path-pairs. In each building block, we represent paths as nodes and interactions (i.e., shared common individuals between two paths) as edges. There are at least two paths in a building block. For each building block, we obtain all acceptable cases for concerned path-pairs. Given a path-triple, it can be decomposed to one or multiple building blocks. Considering a shared path-pair between two building blocks, we use the *natural join* operator from relational algebra to match the acceptable cases for the shared path-pair between two building blocks. In other words, considering the acceptable cases for building blocks as inputs, we use the natural join operator to construct all acceptable cases for a path-triple. Acceptable cases for a path-triple are identified and then used in deriving the path-counting formula for Φ_{abc}.

Then, we summarize all the main procedures used for deriving the path-counting formula for Φ_{abc} in a flowchart shown in Figure 4. The main procedures are also applicable for deriving the path-counting formulas for Φ_{abcd} and $\Phi_{ab,cd}$.

3. Results and Discussion

3.1. Path-Counting Formulas for Three Individuals. We first introduce a systematic method to generate all possible cases

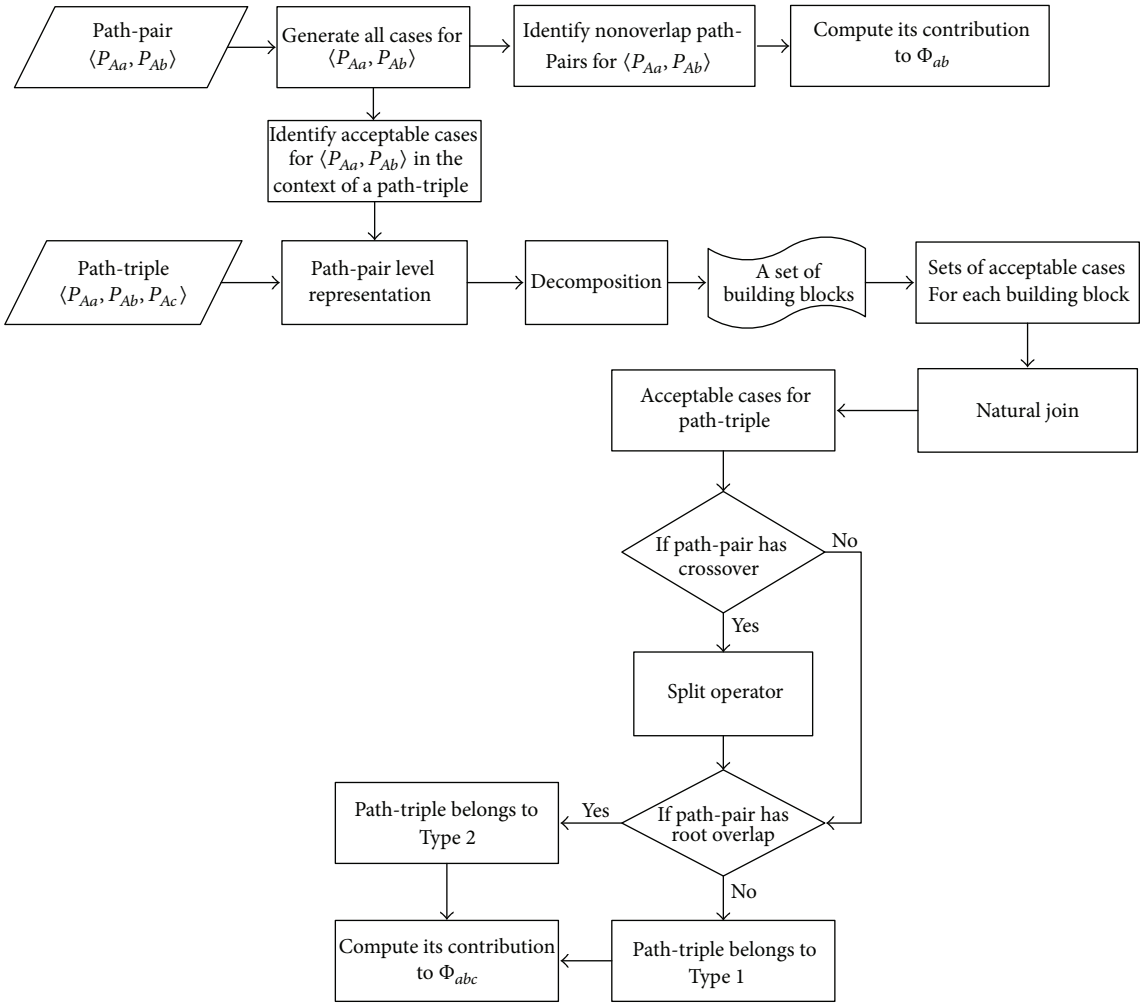


Figure 4: A flowchart for path-counting formula derivation.

for a path-pair. Then we discuss building blocks for path-triples and identify all acceptable cases which are used in deriving the path-counting formula for Φ_{abc}.

3.1.1. Cases for a Path-Pair. Given a path-pair $\langle P_{Aa}, P_{Ab}\rangle$ with $Bi\_C(P_{Aa}, P_{Ab}) \neq NULL$, where A is a common ancestor of a and b and $Bi\_C(P_{Aa}, P_{Ab})$ consists of all common individuals shared between P_{Aa} and P_{Ab}, except A, we introduce three patterns (i.e., *crossover*, 2-*overlap,* and *root 2-overlap*) to generate all possible cases for $\langle P_{Aa}, P_{Ab}\rangle$.

(1) $X(P_{Aa}, P_{Ab})$: P_{Aa} and P_{Ab} share one or multiple crossover individuals.

(2) $T(P_{Aa}, P_{Ab})$: P_{Aa} and P_{Ab} are root 2-overlapping from A, and the root 2-overlap path can have one or multiple 2-overlap individuals.

(3) $Y(P_{Aa}, P_{Ab})$: P_{Aa} and P_{Ab} are overlapping but not from A, and the 2-overlap path can have one or multiple 2-overlap individuals.

Based on the three patterns, $X(P_{Aa}, P_{Ab})$, $T(P_{Aa}, P_{Ab})$, and $Y(P_{Aa}, P_{Ab})$, we use regular expressions to generate all possible cases for the path-pair $\langle P_{Aa}, P_{Ab}\rangle$. For convenience, we drop $\langle P_{Aa}, P_{Ab}\rangle$ and use X, T, and Y instead of patterns $X(P_{Aa}, P_{Ab})$, $T(P_{Aa}, P_{Ab})$, and $Y(P_{Aa}, P_{Ab})$, whenever there is no confusion. When $Bi\_C(P_{Aa}, P_{Ab}) \neq NULL$, the eight cases shown in (7) cover all possible cases for $\langle P_{Aa}, P_{Ab}\rangle$. The completeness of eight cases shown in (7) for $\langle P_{Aa}, P_{Ab}\rangle$ can be proved by induction on the total number of T, X, and Y appearing in $\langle P_{Aa}, P_{Ab}\rangle$. Using the pedigree in Figure 2, Cases 1–3 and Case 6 are illustrated in (8), (9), (10), and (11):

$$\begin{cases} \text{Case 1: } T \\ \text{Case 2: } X^{+} \\ \text{Case 3: } TX^{+} \\ \text{Case 4: } T(X^{+}Y)^{+}, \end{cases}$$
$$\begin{cases} \text{Case 5: } T(X^{+}Y)^{+}X^{+} \\ \text{Case 6: } X^{+}Y \\ \text{Case 7: } X^{+}(YX^{+})^{+} \\ \text{Case 8: } X^{+}(YX^{+})^{+}Y, \end{cases} \tag{7}$$

$$\left.\begin{array}{l} A \longrightarrow s \longrightarrow e \longrightarrow t \longrightarrow a \\ A \longrightarrow s \longrightarrow e \longrightarrow t \longrightarrow b \end{array}\right\} \in T, \tag{8}$$

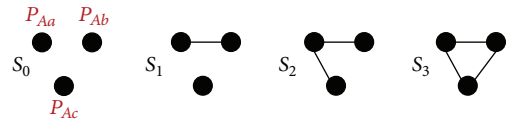


Figure 5: A path-pair level graphical representation of $\langle P_{Aa}, P_{Ab}, P_{Ac}\rangle$.

where $\{s, e, t\}$ are 2-overlap individuals and the overlap path is a root 2-overlap path:

$$\left.\begin{array}{l} A \longrightarrow s \longrightarrow e \longrightarrow t \longrightarrow a \\ A \longrightarrow s \longrightarrow f \longrightarrow t \longrightarrow b \end{array}\right\} \in TX, \tag{9}$$

where s is a 2-overlap individual and the overlap path is a root 2-overlap path; t is a crossover individual:

$$\left.\begin{array}{l} A \longrightarrow s \longrightarrow e \longrightarrow t \longrightarrow a \\ A \longrightarrow d \longrightarrow f \longrightarrow t \longrightarrow b \end{array}\right\} \in X, \tag{10}$$

where t is a crossover individual:

$$\left.\begin{array}{l} A \longrightarrow c \longrightarrow e \longrightarrow t \longrightarrow a \\ A \longrightarrow s \longrightarrow e \longrightarrow t \longrightarrow b \end{array}\right\} \in XY, \tag{11}$$

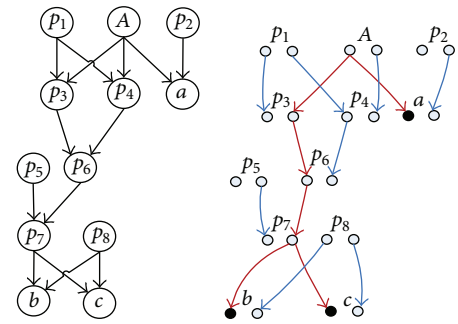
where e is a crossover individual; t is a 2-overlap individual and the overlap path is a 2-overlap path.

3.1.2. Path-Pair Level Graphical Representation of a Path-Triple. Given a path-triple $\langle P_{Aa}, P_{Ab}, P_{Ac}\rangle$, we represent each path as a node. The path-triple can be decomposed to three path-pairs (i.e., $\langle P_{Aa}, P_{Ab}\rangle$, $\langle P_{Aa}, P_{Ac}\rangle$, and $\langle P_{Ab}, P_{Ac}\rangle$). For each path-pair, if the two paths share at least one common individual (i.e., either 2-overlap individual or crossover individual), except A, then there is an edge between the two nodes representing the two paths. Therefore, we obtain four different scenarios S_0–S_3, shown in Figure 5.

In Figure 5, the scenario S_0 has no edges, so it means that $\langle P_{Aa}, P_{Ab}, P_{Ac}\rangle$ consists of three independent paths. In Figure 2, *path-triple*1 is an example of S_0. Next, we introduce a lemma which can assist with identifying the options for the edges in the scenarios S_1–S_3.

Lemma 3. *Given a path-triple $\langle P_{Aa}, P_{Ab}, P_{Ac}\rangle$, consider the three path-pairs $\langle P_{Aa}, P_{Ab}\rangle$, $\langle P_{Aa}, P_{Ac}\rangle$, and $\langle P_{Ab}, P_{Ac}\rangle$, if there is a 2-overlap edge which is represented by Y in regular expression representation of any of the three path-pairs, and then the path-triple $\langle P_{Aa}, P_{Ab}, P_{Ac}\rangle$ has no contribution to Φ_{abc}.*

Proof. In [17], Nadot and Vaysseix proposed, from a genetic and biological point of view, that Φ_{abc} can be evaluated by enumerating all eligible inheritance paths at allele-level starting from a triple common ancestor A to the three individuals a, b, and c.



(a) Pedigree (b) Inheritance paths

Figure 6: Examples of pedigree and inheritance paths.

For the pedigree in Figure 6, let us consider the path-triple $\langle P_{Aa}, P_{Ab}, P_{Ac}\rangle$ listed as follows. $P_{Aa}: A \rightarrow a$; $P_{Ab}: A \rightarrow p_3 \rightarrow p_6 \rightarrow p_7 \rightarrow b$; $P_{Ac}: A \rightarrow p_4 \rightarrow p_6 \rightarrow p_7 \rightarrow c$.

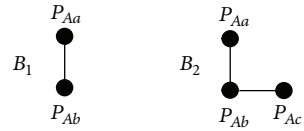
For $\langle P_{Ab}, P_{Ac}\rangle$, p_6 is a crossover individual, p_7 is an overlap individual, and $p_6 \rightarrow p_7$ is a 2-overlap edge repre-sented by Y in regular expression representation (see the definition for Y in Section 3.1.1).

For the individual p_6, let us denote the two alleles at one fixed autosomal locus as g_1 and g_2. At allele-level, only one allele can be passed down from p_6 to p_7. Since p_3 and p_4 are parents of p_6, g_1 is passed down from one parent, and g_2 is passed down from the other parent. It is infeasible to pass down both g_1 and g_2 from p_6 to p_7. In other words, there are no corresponding inheritance paths for the path-triple $\langle P_{Aa}, P_{Ab}, P_{Ac}\rangle$ with a 2-overlap edge between $\langle P_{Ab}, P_{Ac}\rangle$ (i.e., Case 6: XY). Therefore, such kind of path-triples has no contribution to Φ_{abc}. □

Figure 6(b) shows one example of eligible inheritance paths corresponding to a pedigree graph. Each individual is represented by two allele nodes. The eligible inheritance paths in Figure 6(b) consist of red edges only.

Only Case 1, Case 2, and Case 3 do not have Y in the regular expression representation of a path-pair (see (7)); considering the scenarios S_1–S_3 shown in Figure 5, an edge can have three options {Case 1: T; Case 2: X; Case 3: TX}.

3.1.3. Constructing Cases for a Path-Triple. For the scenarios S_1–S_3 in Figure 5, we define two building blocks $\{B_1, B_2\}$ along with some rules in Figure 7 to generate acceptable cases. For B_1, the edge can have three options {Case 1: T; Case 2: X; Case 3: TX}. For B_2, we cannot allow both edges to be root overlap, because if two edges are root overlap, then



For B_1, the edge can have three options {case 1: T, case 2: X, case 3: TX}.
For B_2, there can be at most one edge belonging to root overlap (either T or TX).

Figure 7: Building blocks $\{B_1, B_2\}$ and basic rules.

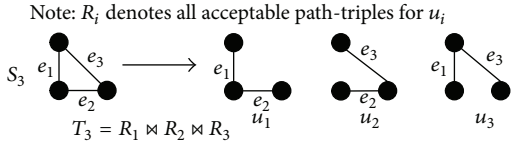


Figure 8: A graphical illustration for obtaining T_3.

P_{Aa} and P_{Ac} must share at least one com-mon individual, except A, which contradicts the fact that P_{Aa} and P_{Ac} have no edge.

Next, we focus on generating all acceptable cases for the scenarios S_1–S_3 in Figure 5, where only S_3 contains more than one building block. In order to leverage the dependency among building blocks, we decompose S_3 to $S_3 = \{u_1 = B_2, u_2 = B_2, u_3 = B_2\}$, shown in Figure 8. For each u_i, we have a set of acceptable path-triples, denoted as R_i.

Considering the dependency among $\{R_1, R_2, R_3\}$, we use the natural join operator, denoted as $\bowtie$, operating on $\{R_1, R_2, R_3\}$ to generate all acceptable cases for S_3. As a result, we obtain $T_3 = R_1 \bowtie R_2 \bowtie R_3$, where T_3 denotes the acceptable cases of the path-triple $\langle P_{Aa}, P_{Ab}, P_{Ac}\rangle$ in the scenario S_3.

For each scenario in Figure 5, we generate all acceptable cases for $\langle P_{Aa}, P_{Ab}, P_{Ac}\rangle$. The scenario S_0 has no edges, and it shows that $\langle P_{Aa}, P_{Ab}, P_{Ac}\rangle$ consists of three independent paths, while, for the other scenarios S_k ($k = 1, 2, 3$), the k edges can have two options:

(1) all k edges belong to *crossover*; or

(2) one edge belongs to *root 2-overlap*; the remaining (k – 1) edges belong to *crossover*.

In summary, acceptable path-triples can have at most one root 2-overlap path, any number of crossover individuals, but zero 2-overlap path.

3.1.4. Splitting Operator. Considering the existence of root 2-overlap path and crossover in acceptable path-triples, we propose a splitting operator to transform a path-triple with crossover individuals to a noncrossover path-triple without changing the contribution from this path-triple to Φ_{abc}. The main purpose of using the splitting operator is to simplify the path-counting formula derivation process. We first use an example in Figure 9 to illustrate how the splitting operator works. In Figure 9, there is a crossover individual s between P_{Aa} and P_{Ab} in the path triple $\langle P_{Aa}, P_{Ab}, P_{Ac}\rangle$ in G_{k+1}. The splitting operator proceeds as follows:

(1) split the node s to two nodes, s_1 and s_2;

(2) transform the edges $s \to a'$ and $s \to b'$ to $s_1 \to a'$ and $s_2 \to b'$, respectively;

(3) add two new edges, $s_2 \to a'$ and $s_1 \to b'$.

Lemma 4. *Given a pedigree graph G_{k+1} having $(k + 1)$ crossover individuals regarding $\langle P_{Aa}, P_{Ab}, P_{Ac}\rangle$ shown in Figure 9, let s denote the lowest crossover individual, where no descendant of s can be a crossover individual among the three paths P_{Aa}, P_{Ab}, and P_{Ac}. After using the splitting operator for the lowest crossover individual s in $G_k + 1$, the number of crossover individuals in G_{k+1} is decreased by 1.*

Proof. The splitting operator only affects the edges from s to a' and b'. If there is a new crossover node appearing, the only possible node is either a' or b'. Assume b' becomes a crossover individual; it means that b' is able to reach a and b from two separate paths. It contradicts the fact that s is the lowest crossover individual between P_{Aa} and P_{Ab}. □

Next, we introduce a canonical graph which results from applying the splitting operator for all crossover individuals. The canonical graph has zero crossover individual.

Definition 5 (Canonical Graph). Given a pedigree graph G having one or more crossover individuals regarding Φ_{abc}, If there exists a graph G' which has no crossover individuals with regards to Φ_{abc} such that

(i) any acceptable path-triple in G has an acceptable path-triple in G' which has the same contribution to Φ_{abc} as the one in G for Φ_{abc};

(ii) any acceptable path-triple in G' has an acceptable path-triple in G which and has the same contribution to Φ_{abc} as the one in G' for Φ_{abc}.

We call G' a *canonical graph* of G regarding Φ_{abc}.

Lemma 6. *For a pedigree graph G having one or more crossover individuals regarding $\langle P_{Aa}, P_{Ab}, P_{Ac}\rangle$, there exists a canonical graph G' for G.*

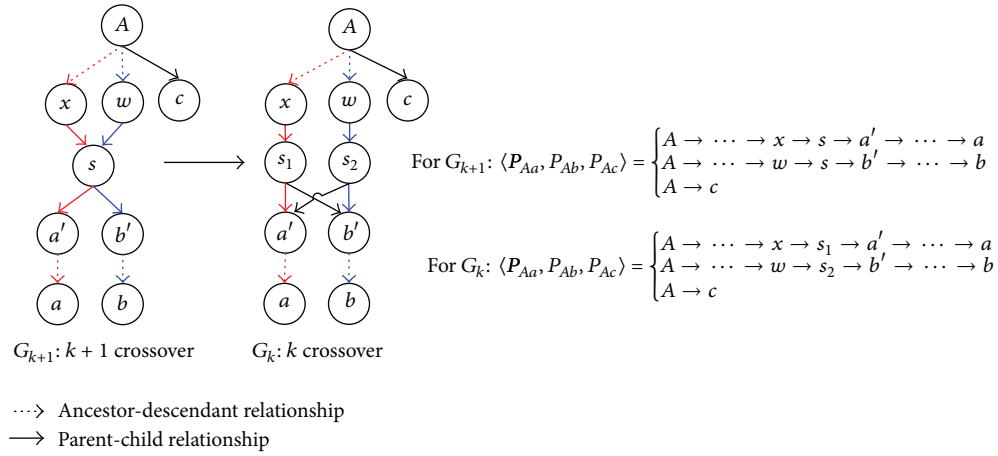


FIGURE 9: Transforming pedigree graph $G_k + 1$ having $k + 1$ crossover to G_k having k crossover.

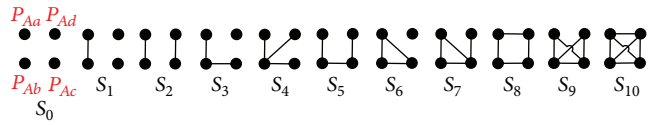


FIGURE 10: A path-pair level graphical representation of $\langle P_{Aa}, P_{Ab}, P_{Ac}, P_{Ad}\rangle$.

Proof (Sketch). The proof is by induction on the number of crossover individuals.

Induction hypothesis: assume that if G has k or less crossovers, there is a canonical graph G' for G.

In the induction step, let G_{k+1} be a graph with $k+1$ crossovers; let s be the lowest crossover between paths P_{Aa} and P_{Ab} in G_{k+1}. We apply the splitting operator on s in G_{k+1} and obtain G_k having k crossovers by Lemma 4. □

3.1.5. Path-Counting Formula for Φ_{abc}. Now, we present the path-counting formula for Φ_{abc}:

$$\Phi_{abc} = \sum_{A}\left(\sum_{\text{Type 1}}\left(\frac{1}{2}\right)^{L_{\text{triple}}}\Phi_{AAA} + \sum_{\text{Type 2}}\left(\frac{1}{2}\right)^{L_{\text{triple}}+1}\Phi_{AA}\right), \tag{12}$$

where $\Phi_{AA} = (1/2)(1 + F_A)$, $\Phi_{AAA} = (1/4)(1 + 3F_A)$, F_A: the inbreeding coefficient of A, A: a triple-common ancestor of a, b, and c, Type 1: $\langle P_{Aa}, P_{Ab}, P_{Ac}\rangle$ has zero root 2-overlap, Type 2: $\langle P_{Aa}, P_{Ab}, P_{Ac}\rangle$ has one root 2-overlap path P_{As} ending at the individual s

$$L_{\text{triple}} = \begin{cases} L_{P_{Aa}} + L_{P_{Ab}} + L_{P_{Ac}} & \text{for Type 1} \\ L_{P_{Aa}} + L_{P_{Ab}} + L_{P_{Ac}} - L_{P_{As}} & \text{for Type 2,} \end{cases} \tag{13}$$

and $L_{P_{Aa}}$: the length of the path P_{Aa} (also applicable for P_{Aa}, P_{Ac}, and P_{As}).

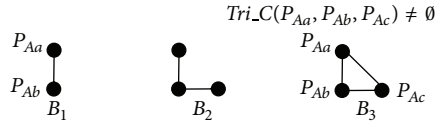
For completeness, the path-counting formula for Φ_{aab} is given in Appendix A; and the correctness proof of the path-counting formula is given in Appendix B.

3.2. Path-Counting Formulas for Four Individuals

3.2.1. Path-Pair Level Graphical Representation of $\langle P_{Aa}, P_{Ab}, P_{Ac}, P_{Ad}\rangle$. Given a path-quad $\langle P_{Aa}, P_{Ab}, P_{Ac}, P_{Ad}\rangle$ and $Quad\_C(P_{Aa}, P_{Ab}, P_{Ac}, P_{Ad}) = \emptyset$, the path-quad can have 11 scenarios S_0–S_{10} shown in Figure 10 where all four paths are considered symmetrically.

In Figure 11, we introduce three building blocks $\{B_1, B_2, B_3\}$. For B_1 and B_2, the rules presented in Figure 7 are also applicable for Figure 11. For B_3, we only consider root overlap, because the crossover individuals can be eliminated by using the splitting operator introduced in Section 3.1.4. Note that for B_3, if $Tri\_C(P_{Aa}, P_{Ab}, P_{Ac}) = \emptyset$, then it is equivalent to the scenario S_3 in Figure 8 Therefore, we only need to consider B_3 when $Tri\_C(P_{Aa}, P_{Ab}, P_{Ac}) \neq \emptyset$.

3.2.2. Building Block-Based Cases Construction for $\langle P_{Aa}, P_{Ab}, P_{Ac}, P_{Ad}\rangle$. For a scenario S_i $(0 \leq i \leq 10)$ in Figure 11, we first decompose S_i to one or multiple building blocks. For a scenario $S_i \in \{S_1, S_3\}$, it has only one building block, and all acceptable cases can be obtained directly. For $S_2 = \{u_1 = B_1, u_2 = B_1\}$, there is no need to consider the conflict between the edges in u_1 and u_2 because u_1 and u_2 are disconnected. Let R_i denote all acceptable cases of the path-pairs in u_i, and let T_i denote all acceptable cases for S_i. Therefore, we obtain $T_2 = R_1 \times R_2$ where $\times$ denotes the Cartesian product operator from relational algebra.



For B_3, all three edges belong to root overlap (i.e., having root 3-overlap).

Figure 11: Building blocks for all scenarios of $\langle P_{Aa}, P_{Ab}, P_{Ac}, P_{Ad}\rangle$.

Table 2: Largest subgraph of a scenario S_i ($4 \le i \le 10$ and $i \neq 6$).

S_i	S_4	S_5	S_7	S_8	S_9	S_{10}
S_j	S_3	S_3	S_6	S_5	S_7	S_9

For $S_6 = \{u_1 = B_3\}$, we obtain $T_6 = R_1$. For $S_i \in \{S_i \mid 4 \le i \le 10 \text{ and } i \neq 6\}$, we define the largest subgraph of S_i based on which we construct T_i.

Definition 7 (Largest Subgraph). Given a scenario S_i ($4 \le i \le 10$ and $i \neq 6$), *the largest subgraph* of S_i, denoted as S_j, is defined as follows:

(1) S_j is a proper subgraph of S_i;

(2) if S_i contains B_3, then S_j must also contain B_3;

(3) no such S_k exists that S_j is a proper subgraph of S_k while S_k is also a proper subgraph of S_i.

For each scenario S_i ($4 \le i \le 10$ and $i \neq 6$), we list the largest subgraph of S_i, denoted as S_j, in Table 2.

For a scenario S_i ($4 \le i \le 10$ and $i \neq 6$), let $\text{Diff}(S_i \setminus S_j)$ denote the set of building blocks in S_i but not in S_j, where S_j is the largest subgraph of S_i. Let $|E_i|$ and $|E_j|$ denote the number of edges in S_i and S_j, respectively. According to Table 2, we can conclude that $|E_i| - |E_j| = 1$. In order to leverage the dependency among building blocks, we consider only B_2 in $\text{Diff}(S_i \setminus S_j)$. For example, $\text{Diff}(S_5 \setminus S_3) = \{B_2\}$. Let T_3 denote all acceptable cases for S_3. And let R_1 denote the set of acceptable cases for $\text{Diff}(S_5 \setminus S_3)$. Then, we can use S_3 and $\text{Diff}(S_5 \setminus S_3)$ to construct all acceptable cases for S_5. Then, we apply this idea for constructing all acceptable cases for each S_i in Table 2.

Given a path-quad $\langle P_{Aa}, P_{Ab}, P_{Ac}, P_{Ad}\rangle$, an acceptable case has the following properties:

(1) if there is one root 3-overlap path, there can be at most one root 2-overlap path;

(2) otherwise, there can be at most two root 2-overlap paths.

3.2.3. Path-Counting Formula for Φ_{abcd}. Now, we present the path-counting formula for Φ_{abcd} as follows:

$$\Phi_{abcd} = \sum_{A} \left(\sum_{\text{Type 1}} \left(\frac{1}{2}\right)^{L_{\text{quad}}} \Phi_{AAAA} + \sum_{\text{Type 2}} \left(\frac{1}{2}\right)^{L_{\text{quad}}+1} \Phi_{AAA} + \sum_{\text{Type 3}} \left(\frac{1}{2}\right)^{L_{\text{quad}}+2} \Phi_{AA} \right), \tag{14}$$

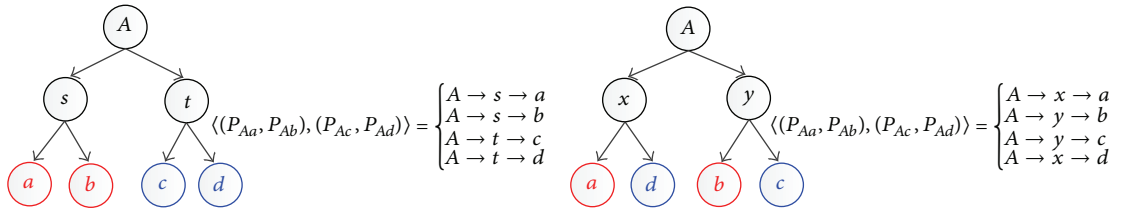
where $\Phi_{AA} = (1/2)(1+F_A)$, $\Phi_{AAA} = (1/4)(1+3F_A)$, $\Phi_{AAAA} = (1/8)(1+7F_A)$, F_A: the inbreeding coefficient of A, A: a quad-common ancestor of a, b, c, and d, Type 1: zero root 2-overlap and zero root 3-overlap path, Type 2: one root 2-overlap path P_{As} ending at s

$$\text{Type 3:} \begin{cases} \text{Case 1: two root 2-overlap paths } P_{As1}, \\ \quad P_{As2} \text{ ending at } s_1 \text{ and } s_2, \text{respectively} \\ \text{Case 2: one root 3-overlap path} \\ \quad P_{At} \text{ ending at } t \\ \text{Case 3: one root 2-overlap path} \\ \quad P_{As}, \text{one root 3-overlap} \\ \quad \text{path } P_{At} \text{ ending at } s \text{ and } t, \\ \quad \text{respectively,} \end{cases}$$

$$L_{\text{quad}} = \begin{cases} L_{P_{Aa}} + L_{P_{Ab}} + L_{P_{Ac}} + L_{P_{Ad}} & \text{for Type 1} \\ L_{P_{Aa}} + L_{P_{Ab}} + L_{P_{Ac}} + L_{P_{Ad}} - L_{P_{As}} & \text{for Type 2} \\ L_{P_{Aa}} + L_{P_{Ab}} + L_{P_{Ac}} + L_{P_{Ad}} - L_{P_{As_1}} - L_{P_{As_2}} & \text{for Case 1} \in \text{Type 3} \\ L_{P_{Aa}} + L_{P_{Ab}} + L_{P_{Ac}} + L_{P_{Ad}} - 2 * L_{P_{At}} & \text{for Case 2} \in \text{Type 3} \\ L_{P_{Aa}} + L_{P_{Ab}} + L_{P_{Ac}} + L_{P_{Ad}} - L_{P_{At}} - L_{P_{As}} & \text{for Case 3} \in \text{Type 3,} \end{cases} \tag{15}$$

and $L_{P_{Aa}}$: the length of the path P_{Aa} (also applicable for P_{Ab}, P_{Ac}, P_{Ad}, etc.).

For completeness, the path-counting formulas for Φ_{aabc} and Φ_{aaab} are presented in Appendix A. The correctness of the path-counting formula for four individuals is proven in Appendix C.



(a) (b)

Figure 12: Examples of 2-pair-path-quads for $\Phi_{ab,cd}$.

3.3. Path-Counting Formulas for Two Pairs of Individuals

3.3.1. Terminology and Definitions

(1) 2-Pair-Path-Pair. It consists of two pairs of path-pairs denoted as $\langle (P_{Sa}, P_{Sb}), (P_{Tc}, P_{Td})\rangle$, where $P_{Sa} \in P(S, a)$, $P_{Sb} \in P(S, b)$, $P_{Tc} \in P(T, c)$, $P_{Td} \in P(T, d)$, S is a common ancestor of a and b, and T is a common ancestor of c and d. If $A = S = T$, then A is a *quad-common ancestor* of a, b, c, and d.

(2) Homo-Overlap and Heter-Overlap Individual. Given two pairs of individuals $\langle a, b\rangle$ and $\langle c, d\rangle$, if $s \in Bi\_C(P_{Aa}, P_{Ab})$ (or $s \in Bi\_C(P_{Ac}, P_{Ad})$, we call s a *homo-overlap individual* when P_{Aa} and P_{Ab} (or P_{Ac} and P_{Ad}) pass through the *same* parent of s. If $r \in Bi\_C(P_{Ai}, P_{Aj})$, where $i \in \{a, b\}$ and $j \in \{c, d\}$, we call r a *heter-overlap individual* when P_{Ai} and P_{Aj} pass through the *same* parent of r.

(3) Root Homo-Overlap and Heter-Overlap Path. Given a 2-pair-path-pair $\langle (P_{Aa}, P_{Ab}), (P_{Ac}, P_{Ad})\rangle$, if s is a homo-overlap individual and the homo-overlap path extends all the way to the quad-common ancestor A, then we call it a *root homo-overlap path*. If r is a heter-overlap individual and the heter-overlap path extends all the way to the quad-common ancestor A, then we call it a *root heter-overlap path*.

Example 8. A is *quad-common ancestor* for a, b, c, and d in Figure 12. For (a), s is a *homo-overlap* individual between P_{Aa} and P_{Ab}.

t is a *homo-overlap* individual between P_{Ac} and P_{Ad}. And, $A \to s$ and $A \to t$ are *root homo-overlap* paths. For (b), x is a *heter-overlap* individual between P_{Aa} and P_{Ad}. y is a *heter-overlap* individual between P_{Ab} and P_{Ac}. And $A \to x$ and $A \to y$ are *root heter-overlap* paths.

3.3.2. Path-Counting Formula for $\Phi_{ab,cd}$.

Now, we present a path-pair level graphical representation for $\langle (P_{Aa}, P_{Ab}), (P_{Ac}, P_{Ad})\rangle$ shown in Figure 13. The options for an edge can be $\{T, X, TX\}$. (Refer to Section 3.1.1 for definitions of T, X, and TX). Based on the different types of $\langle P_{Aa}, P_{Ab}, P_{Ac}, P_{Ad}\rangle$ presented in (14), all cases for $\langle (P_{Aa}, P_{Ab}), (P_{Ac}, P_{Ad})\rangle$ are summarized in Table 3, where h is the last individual of a root homo-overlap path P_{Ah} (i.e., the path P_{Ah} ending at h) and r_1 and r_2 are the last individuals of root heter-overlap paths P_{Ar1} and P_{Ar2}, respectively.

Given a pedigree graph having one or multiple progenitors $\{p_i \mid i > 0\}$, we define that the generation of a progenitor p_i is 0, denoted as $\text{gen}(p_i) = 0$. If an individual a has only one parent p, then we define $\text{gen}(a) = \text{gen}(p) + 1$. If an individual a has two parents f and m, we define $\text{gen}(a) = \text{MAX}\{\text{gen}(f), \text{gen}(m)\} + 1$.

Table 3: A summary of all cases for $\langle (P_{Aa}, P_{Ab}), (P_{Ac}, P_{Ad})\rangle$.

$\langle P_{Aa}, P_{Ab}, P_{Ac}, P_{Ad}\rangle$	$\langle (P_{Aa}, P_{Ab}), (P_{Ac}, P_{Ad})\rangle$
Zero root 2-overlap and zero root 3-overlap	Zero root homo-overlap and zero root heter-overlap
One root 2-overlap path	One root homo-overlap and zero root heter-overlap Zero root homo-overlap and one root heter-overlap
Two root 2-overlap paths	Two root homo-overlaps and zero root heter-overlap Zero root homo-overlap and two root heter-overlaps
One root 3-overlap path	One root homo-overlap and two root heter-overlaps, and $h = r_1 = r_2$
One root 2-overlap and one root 3-overlap	One root homo-overlap and two root heter-overlaps, and $r_1 = r_2 \neq h$ One root homo-overlap and two root heter-overlaps, and $h = r_1 \neq r_2$

The path-counting formula for $\Phi_{ab,cd}$ is as follows:

$$\Phi_{ab,cd} = \sum_{A} \left(\sum_{\text{Type 1}} \left(\frac{1}{2}\right)^{L_{2\text{-pair}}} \Phi_{AAA} + \sum_{\text{Type 2}} \left(\frac{1}{2}\right)^{L_{2\text{-pair}}+1} \Phi_{AAA} + \sum_{\text{Type 3}} \left(\frac{1}{2}\right)^{L_{2\text{-pair}}+2} \Phi_{AA} + \sum_{\text{Type 4}} \left(\frac{1}{2}\right)^{L_{2\text{-pair}}+1} \Phi_{AA} \right) + \sum_{(S,T)\in\text{Type 5}} \left(\frac{1}{2}\right)^{L_{\langle P_{Sa}, P_{Sb}\rangle}+L_{\langle P_{Tc}, P_{Td}\rangle}+1} \Phi_{BB}, \tag{16}$$

where A: a quad-common ancestor of a, b, c, and d, S: a common ancestor of a and b, and T: a common ancestor of c and d. For $\langle (P_{Aa}, P_{Ab}), (P_{Ac}, P_{Ad})\rangle$ ($S = T = A$), there are four types (i.e., Type 1 to Type 4).

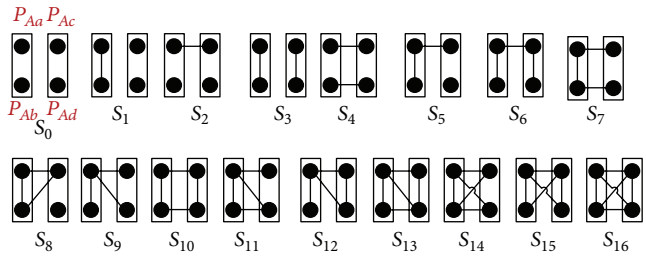


Figure 13: Scenarios of $\langle (P_{Aa}, P_{Ab}), (P_{Ac}, P_{Ad}) \rangle$ at path-pair level.

Type 1: zero root homo-overlap and zero root heter-overlap.

Type 2: zero root homo-overlap and one root heter-overlap P_{Ar} ending at r,

$$\text{Type 3: } \begin{cases} \text{zero root homo-overlap and two root} \\ \quad \text{heter-overlap } P_{Ar1} \text{ and } P_{Ar2} \text{ ending at} \\ \quad r_1 \text{ and } r_2, \text{respectively,} \\ \text{one root homo-overlap } P_{Ah} \text{ ending at } h \\ \quad \text{and two root heter-overlap } P_{Ar1} \text{ and } P_{Ar2} \\ \quad \text{ending at } r_1 \text{ and } r_2, \text{and } r_1 \neq r_2. \end{cases} \tag{17}$$

Type 4: one root homo-overlap P_{Ah} ending at h and two root heter-overlap ending at r_1 and r_2, and $h = r_1 = r_2$. For $\langle (P_{Sa}, P_{Sb}), (P_{Tc}, P_{Td}) \rangle$ $(S \neq T)$, there is one type (i.e., Type 5).

Type 5: $\langle P_{Sa}, P_{Sb} \rangle$ has zero overlap individual, $\langle P_{Tc}, P_{Td} \rangle$ has zero overlap individual.

At most one path-pair (either $\langle P_{Sa}, P_{Sb} \rangle$ or $\langle P_{Tc}, P_{Td} \rangle$) can have crossover individuals.

Between a path from $\langle P_{Sa}, P_{Sb} \rangle$ and a path from $\langle P_{Tc}, P_{Td} \rangle$, there are no overlap individuals, but there can be crossover individuals, x, where $x \neq S$ and $x \neq T$:

$$B = \begin{cases} S & \text{when gen}(S) < \text{gen}(T) \\ S & \text{when gen}(S) = \text{gen}(T) \\ & \quad \text{and } T \text{ has two parents} \\ T & \text{otherwise,} \end{cases}$$

$$L_{\text{2-pair}} = \begin{cases} L_{P_{Aa}} + L_{P_{Ab}} & \\ \quad + L_{P_{Ac}} + L_{P_{Ad}} & \text{for Type 1} \\ L_{P_{Aa}} + L_{P_{Ab}} + L_{P_{Ac}} & \\ \quad + L_{P_{Ad}} - L_{P_{Ar}} & \text{for Type 2} \\ L_{P_{Aa}} + L_{P_{Ab}} + L_{P_{Ac}} & \\ \quad + L_{P_{Ad}} - L_{P_{Ar1}} - L_{P_{Ar2}} & \text{for Type 3} \\ L_{P_{Aa}} + L_{P_{Ab}} + L_{P_{Ac}} & \\ \quad + L_{P_{Ad}} - 2 * L_{P_{Ah}} & \text{for Type 4,} \end{cases} \tag{18}$$

$$L_{\langle P_{Sa}, P_{Sb} \rangle} = L_{P_{Sa}} + L_{P_{Sb}} \quad \text{for Type 5,}$$

$$L_{\langle P_{Tc}, P_{Td} \rangle} = L_{P_{Tc}} + L_{P_{Td}} \quad \text{for Type 5.}$$

Note that if $\langle a, b \rangle$ and $\langle c, d \rangle$ have zero quad-common ancestors, we have the following formula for $\Phi_{ab,cd}$:

$$\Phi_{ab,cd} = \sum_{(S,T) \in \text{Type 6}} \left(\frac{1}{2} \right)^{L_{\langle P_{Sa}, P_{Sb} \rangle} + L_{\langle P_{Tc}, P_{Td} \rangle}} \Phi_{SS} * \Phi_{TT}. \tag{19}$$

Type 6: $\langle P_{Sa}, P_{Sb} \rangle$ is a nonoverlapping path-pair and $\langle P_{Tc}, P_{Td} \rangle$ is a nonoverlapping path-pair. Between a path from $\langle P_{Sa}, P_{Sb} \rangle$ and a path from $\langle P_{Tc}, P_{Td} \rangle$, there are no overlap individuals, but there can be crossover individuals.

$L_{\langle P_{Sa}, P_{Sb} \rangle}$ and $L_{\langle P_{Tc}, P_{Td} \rangle}$ are defined as in Type 5.

The correctness of the path-counting formula for $\Phi_{ab.cd}$ is proven in Appendix C. For completeness, please refer to [18] for the path-counting formulas for $\Phi_{aa,bc}$, $\Phi_{ab,ac}$, $\Phi_{ab,ab}$, and $\Phi_{aa,ab}$.

3.4. Experimental Results. In this section, we show the efficiency of our path-counting method using NodeCodes for condensed identity coefficients by making comparisons with the performance of a recursive method used in [10]. We implemented two methods: (1) using recursive formulas to compute each required kinship coefficient and generalized kinship coefficient; (2) using path-counting method coupled with NodeCodes to compute each required kinship coefficient and generalized kinship coefficient independently. We refer to the first method as *Recursive*, the second method as *NodeCodes*. For completeness, please refer to [18] for the details of the NodeCodes-based method.

Nodecodes of a node is a set of labels each representing a path to the node from its ancestors. Given a pedigree graph, let r be the progenitor (i.e., the node with 0 in-degree). (For simplicity, we assume there is one progenitor, r, as the ancestor of all individuals in the pedigree. Otherwise, a virtual node r can be added to the pedigree graph and all progenitors can be made children of r.) For each node u in the graph, the set of NodeCodes of u, denoted as NC(u), are assigned using a breadth-first-search traversal starting from r as follows.

(1) If u is r then NC(r) contains only one element: the empty string.

(2) Otherwise, let u be a node with NC(u), and $v_0, v_1, \ldots, v_k$ be u's children in sibling order; then for each x in NC(u), a code xi^* is added to NC(v_i), where $0 \leq i \leq k$, and * indicates the gender of the individual represented by node v_i.

Computations of kinship coefficients for two individuals and generalized kinship coefficients for three individuals presented in [11, 12, 14, 15] are using NodeCodes. The NodeCodes-based computation schemes can also be applied for the generalized kinship coefficients for four individuals and two pairs of individuals. For completeness, please refer to [18] for the details using NodeCodes to compute the generalized kinship coefficients for four individuals and two pairs of individuals based on our proposed path-counting formulas in Sections 3.2 and 3.3.

In order to test the scalability of our approach for calculating condensed identity coefficients on large pedigrees, we used a population simulator implemented in [11] to generate arbitrarily large pedigrees. The population simulator is based on the algorithm for generating populations with overlapping generations in Chapter 4 of [19] along with the parameters given in Appendix B of [20] to model the relatively isolated Finnish Kainuu subpopulation and its growth during the years 1500–2000. An overview of the generation algorithm was presented in [11, 12, 14]. The parameters include starting/ending year, initial population size, initial age distribution, marriage probability, maximum age at pregnancy, expected number of children by time period, immigration rate, and probability of death by time period and age group.

We examine the performance of condensed identity coefficients using twelve synthetic pedigrees which range from 75 individuals to 195,197 individuals. The smallest pedigree spans 3 generations, and the largest pedigree spans 19 generations. We analyzed the effects of pedigree size and the depth of individuals in the pedigree (the longest path between the individual and a progenitor) on the computation efficiency improvement.

In the first experiment, 300 random pairs were selected from each of our 12 synthetic pedigrees. Figure 14 shows computation efficiency improvement for each pedigree. As can be seen, the improvement of *NodeCodes* over *Recursive* grew increasingly larger as the pedigree size increased, from a comparable amount of 26.83% on the smallest pedigree to 94.75% on the largest pedigree. It also shows that path-counting method coupled with NodeCodes can scale very well on large pedigrees in terms of computing condensed identity coefficients.

In our next experiment, we examined the effect of the depth of the individual in the pedigree on the query time. For each depth, we generated 300 random pairs from the largest synthetic pedigree.

Figure 15 shows the effect of depth on the computation efficiency improvement. We can see the improvement of *NodeCodes* over *Recursive*, ranging from 86.48% to 91.30%.

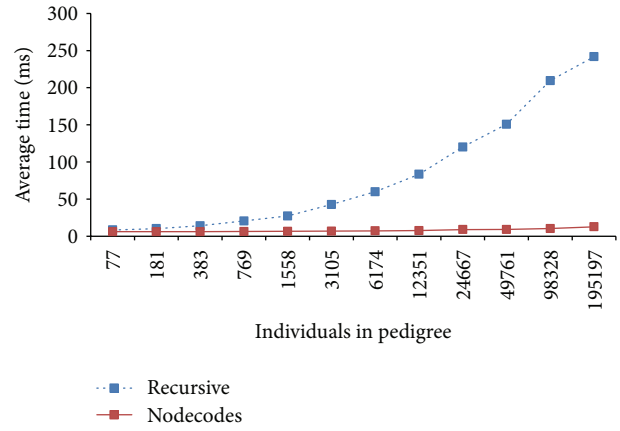


Figure 14: The effect of pedigree size on computation efficiency improvement.

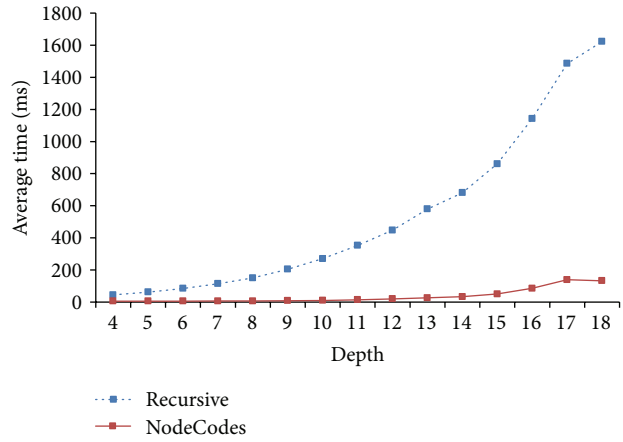


Figure 15: The effect of depth on computation efficiency improvement.

4. Conclusion

We have introduced a framework for generalizing Wright's path-counting formula for more than two individuals. Aiming at efficiently computing condensed identity coefficients, we proposed path-counting formulas (PCF) for all generalized kinship coefficients for which are sufficient for expressing condensed identity coefficients by a linear combination. We also perform experiments to compare the efficiency of our method with the recursive method for computing condensed identity coefficients on large pedigrees. Our future work includes (i) further improvements on condensed identify coefficients computation by collectively calculating the set of generalized kinship coefficients to avoid redundant computations, and (ii) experimental results for using PCF in conjunction with encoding schemes (e.g., compact path-encoding schemes [13]) for computing condensed identity coefficients on very large pedigrees.

Appendices

A. Path-Counting Formulas of Special Cases

A.1. Path-Counting Formula for Φ_{aab}. For $\langle P_{Aa1}, P_{Aa2}\rangle$, we introduce a special case, where P_{Aa1} and P_{Aa2} are *mergeable*.

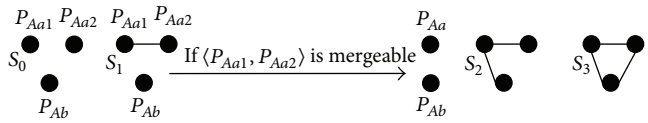


FIGURE 16: A path-pair level graphical representation of $\langle P_{Aa1}, P_{Aa2}, P_{Ab}\rangle$.

Definition A.1 (Mergeable Path-Pair). A path-pair $\langle P_{Aa1}, P_{Aa2}\rangle$ is *mergeable* if and only if the two paths P_{Aa1} and P_{Aa2} are completely identical.

Next, we present a graphical representation of $\langle P_{Aa1}, P_{Aa2}, P_{Ab}\rangle$ in Figure 16.

Lemma A.2. *For S_2 and S_3 in Figure 16, $\langle P_{Aa1}, P_{Aa2}\rangle$ cannot be a mergeable path-pair.*

Proof. For S_2 and S_3, if $\langle P_{Aa1}, P_{Aa2}\rangle$ is mergeable, then any common individual s between P_{Aa1} and P_{Ab} is also a shared individual between P_{Aa2} and P_{Ab}. It means $s \in Tri\_C(P_{Aa1}, P_{Aa2}, P_{Ab})$ which contradicts the fact that $Tri\_C(P_{Aa1}, P_{Aa2}, P_{Ab}) = \emptyset$.

Considering all three scenarios in Figure 16, only S_1 can have a mergeable path-pair $\langle P_{Aa1}, P_{Aa2}\rangle$ by Lemma A.2. Now, we present our path-counting formula for Φ_{aab} where a is not an ancestor of b:

$$\Phi_{aab} = \sum_{A}\left(\sum_{\text{Type 1}}\left(\frac{1}{2}\right)^{L_{\text{triple}}-1}\Phi_{AAA} + \sum_{\text{Type 2}}\left(\frac{1}{2}\right)^{L_{\text{triple}}}\Phi_{AA} + \sum_{\text{Type 3}}\left(\frac{1}{2}\right)^{L_{\langle P_{Aa},P_{Ab}\rangle}+1}\Phi_{AA}\right), \tag{A.1}$$

where A: a common ancestor of a and b.

When $\langle P_{Aa1}, P_{Aa2}\rangle$ is not mergeable,

Type 1: $\langle P_{Aa1}, P_{Aa2}, P_{Ab}\rangle$ has no root 2-overlap.

Type 2: $\langle P_{Aa1}, P_{Aa2}, P_{Ab}\rangle$ has one root 2-overlap path P_{As} ending at the individual s.

When $\langle P_{Aa1}, P_{Aa2}\rangle$ is mergeable,

Type 3: $\langle P_{Aa}, P_{Ab}\rangle$ is a nonoverlapping path-pair

$$L_{\text{triple}} = \begin{cases} L_{P_{Aa1}} + L_{P_{Aa2}} + L_{P_{Ab}} & \text{for Type 1} \\ L_{P_{Aa1}} + L_{P_{Aa2}} + L_{P_{Ab}} - L_{P_{As}} & \text{for Type 2,} \end{cases} \tag{A.2}$$

$$L_{\langle P_{Aa},P_{Ab}\rangle} = L_{P_{Aa}} + L_{P_{Ab}} \quad \text{for Type 3.}$$

For the sake of completeness, if a is an ancestor of b, there is no recursive formula for Φ_{aab} in [10], but we can use either the recursive formula for Φ_{abc} or the path-counting formula for Φ_{abc} to compute Φ_{a1a2b}. □

A.2. Path-Counting Formula for Φ_{aabc}. Given a path-quad $\langle P_{Aa1}, P_{Aa2}, P_{Ab}, P_{Ac}\rangle$, if $\langle P_{Aa1}, P_{Aa2}\rangle$ is not mergeable, then we process the path-quad as equivalent to $\langle P_{Aa}, P_{Ab}, P_{Ac}, P_{Ad}\rangle$. If $\langle P_{Aa1}, P_{Aa2}\rangle$ is mergeable, the path-quad $\langle P_{Aa1}, P_{Aa2}, P_{Ab}, P_{Ac}\rangle$ can be condensed to scenarios for $\langle P_{Aa}, P_{Ab}, P_{Ac}\rangle$.

Now, we present a path-counting formula for Φ_{aabc} where a is not an ancestor of b and c as follows:

$$\Phi_{aabc} = \sum_{A}\left(\sum_{\text{Type 1}}\left(\frac{1}{2}\right)^{L_{\text{quad}}-1}\Phi_{AAAA} + \sum_{\text{Type 2}}\left(\frac{1}{2}\right)^{L_{\text{quad}}}\Phi_{\text{AAA}} + \sum_{\text{Type 3}}\left(\frac{1}{2}\right)^{L_{\text{quad}}+1}\Phi_{AA}\right) + \sum_{A}\left(\sum_{\text{Type 4}}\left(\frac{1}{2}\right)^{L_{\text{triple}}+1}\Phi_{AAA} + \sum_{\text{Type 5}}\left(\frac{1}{2}\right)^{L_{\text{triple}}+2}\Phi_{AA}\right), \tag{A.3}$$

where A: a quad-common ancestor of a, b, c, and d.

When $\langle P_{Aa1}, P_{Aa2}\rangle$ is not mergeable,

Type 1: zero root 2-overlap and zero root 3-overlap path;

Type 2: one root 2-overlap path P_{As} ending at s

$$\text{Type 3:} \begin{cases} \text{Case 1: two root 2-overlap paths } P_{As1} \\ \quad \text{and } P_{As2} \text{ ending at } s_1 \text{ and } s_2\text{, respectively} \\ \text{Case 2: one root 3-overlap path } P_{At} \\ \quad \text{ending at } t \\ \text{Case 3: one root 2-overlap} \\ \quad \text{and one root 3-overlap paths} \\ \quad P_{As} \text{ and } P_{At} \text{ ending at } s \text{ and } t, \\ \quad \text{respectively.} \end{cases} \tag{A.4}$$

When $\langle P_{Aa1}, P_{Aa2}\rangle$ is mergeable,

Type 4: $\langle P_{Aa}, P_{Ab}, P_{Ac}\rangle$ has zero root 2-overlap path;

Type 5: $\langle P_{Aa}, P_{Ab}, P_{Ac}\rangle$ has one root 2-overlap path P_{As} ending at s

$$L_{\text{quad}} = \begin{cases} L_{P_{Aa1}} + L_{P_{Aa2}} + L_{P_{Ab}} + L_{P_{Ac}} & \text{for Type 1} \\ L_{P_{Aa1}} + L_{P_{Aa2}} + L_{P_{Ab}} + L_{P_{Ac}} - L_{P_{As}} & \text{for Type 2} \\ L_{P_{Aa1}} + L_{P_{Aa2}} + L_{P_{Ab}} + L_{P_{Ac}} - L_{P_{As_1}} - L_{P_{As_2}} & \text{for Case 1} \in \text{Type 3} \\ L_{P_{Aa1}} + L_{P_{Aa2}} + L_{P_{Ab}} + L_{P_{Ac}} - L_{P_{At}} & \text{for Case 2} \in \text{Type 3} \\ L_{P_{Aa1}} + L_{P_{Aa2}} + L_{P_{Ab}} + L_{P_{Ac}} - L_{P_{At}} - L_{P_{As}} & \text{for Case 3} \in \text{Type 3,} \end{cases}$$

$$L_{\text{triple}} = \begin{cases} L_{P_{Aa}} + L_{P_{Ab}} + L_{P_{Ac}} & \text{for Type 4} \\ L_{P_{Aa}} + L_{P_{Ab}} + L_{P_{Ac}} - L_{P_{As}} & \text{for Type 5.} \end{cases} \tag{A.5}$$

Note that if a is an ancestor of either b or c, or both of them, then the path-counting formula of Φ_{abcd} is applicable to compute Φ_{a1a2bc}.

A.3. Path-Counting Formula for Φ_{aaab}. A special case of $\langle P_{Aa_1}, P_{Aa_2}, P_{Aa_3}\rangle$ for $\langle P_{Aa_1}, P_{Aa_2}, P_{Aa_3}, P_{Ab}\rangle$ is introduced when $\langle P_{Aa_1}, P_{Aa_2}, P_{Aa_3}\rangle$ is mergeable. With the existence of a mergeable path-triple, $\langle P_{Aa_1}, P_{Aa_2}, P_{Aa_3}, P_{Ab}\rangle$ can be condensed to $\langle P_{Aa}, P_{Ab}\rangle$.

Definition A.3 (Mergeable Path-Triple). Given three paths P_{Aa1}, P_{Aa2}, and P_{Aa3}, they are *mergeable* if and only if they are completely identical.

Lemma A.4. *Given a path-quad $\langle P_{Aa_1}, P_{Aa_2}, P_{Aa_3}, P_{Ab}\rangle$, there must be at least one mergeable path-pair among $\langle P_{Aa_1}, P_{Aa_2}\rangle$, $\langle P_{Aa_1}, P_{Aa_3}\rangle$, $\langle P_{Aa_2}, P_{Aa_3}\rangle$.*

Proof. For an individual a with two parents f and m, the paternal allele of the individual a is transmitted from f and the maternal allele is transmitted from m. At allele level, only two descent paths starting from an ancestor are allowed. For a path-quad $\langle P_{Aa_1}, P_{Aa_2}, P_{Aa_3}, P_{Ab}\rangle$, there must be at least one mergeable path-pair among $\langle P_{Aa_1}, P_{Aa_2}\rangle$, $\langle P_{Aa_1}, P_{Aa_3}\rangle$, and $\langle P_{Aa_2}, P_{Aa_3}\rangle$. □

For simplicity, we treat $\langle P_{Aa1}, P_{Aa2}\rangle$ as a default mergeable path-pair.

Now, we present the path-counting formula for Φ_{aaab} where a is not an ancestor of b as follows:

$$\Phi_{aaab} = \sum_{A}\left(\frac{3}{2}\left(\sum_{\text{Type 1}}\left(\frac{1}{2}\right)^{L_{\text{triple}}-1}\Phi_{AAA} + \sum_{\text{Type 2}}\left(\frac{1}{2}\right)^{L_{\text{triple}}}\Phi_{AA}\right) + \sum_{\text{Type 3}}\left(\frac{1}{2}\right)^{L_{\text{pair}}+2}\Phi_{AA}\right), \tag{A.6}$$

where A: a common ancestor of a and b.

When there is only one mergeable path-pair (let us consider $\langle P_{Aa1}, P_{Aa2}\rangle$ as the mergeable path-pair),

Type 1: $\langle P_{Aa1}, P_{Aa3}, P_{Ab}\rangle$ has zero root 2-overlap path,

Type 2: $\langle P_{Aa1}, P_{Aa3}, P_{Ab}\rangle$ has one root 2-overlap path P_{As} ending at s.

When $\langle P_{Aa1}, P_{Aa2}, P_{Aa3}\rangle$ is mergeable,

Type 3: $\langle P_{Aa}, P_{Ab}\rangle$ is nonoverlapping

$$L_{\text{triple}} = \begin{cases} L_{P_{Aa1}} + L_{P_{Aa3}} + L_{P_{Ab}} & \text{for Type 1} \\ L_{P_{Aa1}} + L_{P_{Aa3}} + L_{P_{Ab}} - L_{P_{As}} & \text{for Type 2,} \end{cases} \tag{A.7}$$

$$L_{\text{pair}} = L_{P_{Aa}} + L_{P_{Ab}} \quad \text{for Type 3.}$$

Note that if a is an ancestor of b, we treat $\Phi_{aaab} = \Phi_{a1a2a3b}$. Then, we apply the path-counting formula for Φ_{abcd} to compute $\Phi_{a1a2a3b}$.

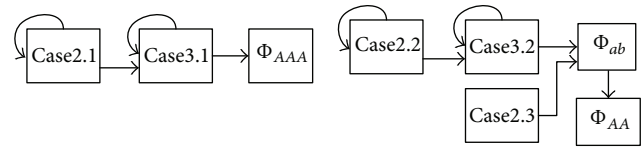


Figure 17: Dependency graph for different cases regarding Φ_{abc} and Φ_{aab}.

B. Proof for Path-Counting Formulas of Three Individuals

We first demonstrate that, for one triple-common ancestor A, the path-counting computation of Φ_{abc} is equivalent to the computation using recursive formulas. Then, we prove the correctness of the path-counting computation for multiple triple-common ancestors.

B.1. One Triple-Common Ancestor. Considering the different types of path-triples starting from a triple-common ancestor A in a pedigree graph G contributing to Φ_{abc} and Φ_{aab}, G can have 5 different cases:

$$\left.\begin{array}{l}\text{Case 2.1: } G \text{ does not have any path-triples } \langle P_{Aa_1}, P_{Aa_2}, P_{Ab}\rangle \text{ with root overlap} \\ \text{Case 2.2: } G \text{ has path-triples } \langle P_{Aa_1}, P_{Aa_2}, P_{Ab}\rangle \text{ with root overlap} \\ \text{Case 2.3: } G \text{ has path-triples } \langle P_{Aa_1}, P_{Aa_2}, P_{Ab}\rangle \text{ having mergeable path-pair } \langle P_{Aa_1}, P_{Aa_2}\rangle\end{array}\right\} \Longleftarrow \Phi_{aab},$$

$$\left.\begin{array}{l}\text{Case 3.1: } G \text{ does not have any path-triples } \langle P_{Aa}, P_{Ab}, P_{Ac}\rangle \text{ with root overlap} \\ \text{Case 3.2: } G \text{ has path-triples } \langle P_{Aa}, P_{Ab}, P_{Ac}\rangle \text{ with root overlap}\end{array}\right\} \Longleftarrow \Phi_{abc}. \tag{B.1}$$

Based on the 5 cases from Case 2.1 to Case 3.2, we first construct a dependency graph shown in Figure 17, consistent with the recursive formulas (3), (4), and (5) for the generalized kinship coefficients for three individuals.

Then, we take the following steps to prove the correctness of the path-counting formulas (12) and (A.1):

(i) for Φ_{ab}, the correctness of the path-counting formula (i.e., Wright's formula) is proven in [21]. For Case 2.1 and Case 2.2, the correctness is proven based on the correctness of Cases 3.1 and 3.2;

(ii) for Case 2.3, it has no cycle but only depends on Φ_{ab}. Thus, we prove the correctness of Case 2.3 by transforming the case to Φ_{ab};

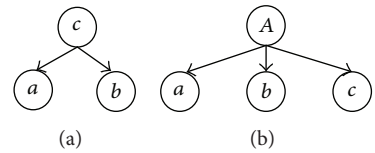


Figure 18: (a) c is a parent of a and b; (b) no individual is a parent of another.

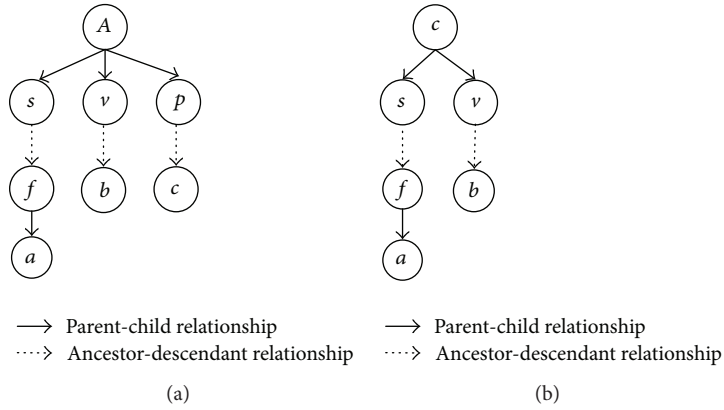


Figure 19: (a) No individual is a parent of another; (b) c is an ancestor of a and b.

(iii) for Cases 3.1 and 3.2, the correctness is proven by induction on the number of edges, n, in the pedigree graph G.

B.1.1. Correctness Proof for Case 3.1

Case 3.1. For Φ_{abc}, G does not have any path triples $\langle P_{Aa}, P_{Ab}, P_{Ac}\rangle$ with root overlap.

Proof (Basis). There are two basic scenarios: (i) one individual is a parent of another; (ii) no individual is a parent of another, among a, b, and c.

Using the recursive formula (3) to compute Φ_{abc}, for Figure 18(a), $\Phi_{abc} = (1/2)\Phi_{cbc} = (1/2)^2\Phi_{ccc}$; for Figure 18(b), $\Phi_{abc} = (1/2)\Phi_{Abc} = (1/2)^2\Phi_{AAc} = (1/2)^3\Phi_{AAA}$.

Using the path-counting formula (12), if a path-triple $\langle P_{Aa}, P_{Ab}, P_{Ac}\rangle$ has no root overlap (i.e., Type 1), then the contribution of $\langle P_{Aa}, P_{Ab}, P_{Ac}\rangle$ to Φ_{abc} can be computed as follows: $\sum_{\text{Type 1}}(1/2)^{L_{\langle P_{Aa},P_{Ab},P_{Ac}\rangle}}\Phi_{AAA}$, where $L_{\langle P_{Aa},P_{Ab},P_{Ac}\rangle} = L_{P_{Aa}} + L_{P_{Ab}} + L_{P_{Ac}}$.

For Figure 18(a), c is the only triple-common ancestor and we obtain $\Phi_{abc} = (1/2)^{L_{\langle P_{ca},P_{cb},P_{cc}\rangle}}\Phi_{ccc} = (1/2)^2\Phi_{ccc}$; for Figure 18(b), we obtain $\Phi_{abc} = (1/2)^{L_{\langle P_{Aa},P_{Ab},P_{Ac}\rangle}}\Phi_{AAA} = (1/2)^3\Phi_{AAA}$.

Induction Step. Let n denote the number of edges in G. Assume true for $n \le k$, where $k \ge 2$. Then, we show it is true for $n = k + 1$.

For Figures 19(a) and 19(b), among a, b, and c, let a be the individual having the longest path starting from their triple-common ancestor in the pedigree graph G with $(k+1)$ edges. If we remove the node a and cut the edge $f \rightarrow a$ from G, then the new graph G^* has k edges. In terms of computing Φ_{fbc}, G^* satisfies the condition for induction hypothesis.

For Figure 19(a), $\Phi_{fbc} = \sum_{\text{Type 1}}(1/2)^{L_{\langle P_{Af},P_{Ab},P_{Ac}\rangle}}\Phi_{AAA}$. Based on the recursive formula (3), $\Phi_{abc} = (1/2)(\Phi_{fbc}+\Phi_{mbc})$ where f and m are parents of a. In G, a only has one parent f; thus, it indicates $\Phi_{mbc} = 0$. Then, we can plug-in the path-counting formula for Φ_{fbc} to obtain

$$\begin{aligned}\Phi_{abc} &= \frac{1}{2}\Phi_{fbc}\\ &= \frac{1}{2} * \sum_{\text{Type 1}}\left(\frac{1}{2}\right)^{L_{\langle P_{Af},P_{Ab},P_{Ac}\rangle}}\Phi_{AAA}\\ &= \sum_{\text{Type 1}}\left(\frac{1}{2}\right)^{L_{\langle P_{Af},P_{Ab},P_{Ac}\rangle}+1}\Phi_{AAA} \qquad \text{(B.2)}\\ &\because L_{\langle P_{Aa},P_{Ab},P_{Ac}\rangle} = L_{\langle P_{Af},P_{Ab},P_{Ac}\rangle} + 1\\ &\therefore \Phi_{abc} = \sum_{\text{Type 1}}\left(\frac{1}{2}\right)^{L_{\langle P_{Aa},P_{Ab},P_{Ac}\rangle}}\Phi_{AAA}.\end{aligned}$$

Similarly, for Figure 19(b), we obtain $\Phi_{abc} = \sum_{\text{Type 1}}(1/2)^{L_{\langle P_{cf},P_{cb},P_{cc}\rangle}+1}\Phi_{ccc} = \sum_{\text{Type 1}}(1/2)^{L_{\langle P_{ca},P_{cb},P_{cc}\rangle}}\Phi_{ccc}$.

Thus, it is true for $n = k + 1$. □

B.1.2. Correctness Proof for Case 3.2

Case 3.2. For Φ_{abc}, G has path triples $\langle P_{Aa}, P_{Ab}, P_{Ac}\rangle$ with root overlap.

Proof (Basis). There are three basic scenarios: (i) there are two individuals who are parents of another; (ii) there is only one individual who is parent of another; (iii) there is no individual who is a parent of another, among a, b, and c.

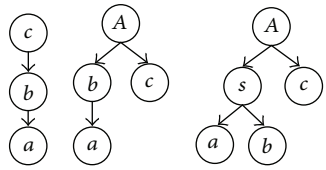


FIGURE 20: (a) b is a parent of a, and c is a parent of b; (b) b is a parent of a; (c) no individual who is a parent of another.

Using the recursive formula (3) to compute Φ_{abc}: in Figure 20, for Figure 20(a), $\Phi_{abc} = (1/2)\Phi_{bbc} = (1/2)^2\Phi_{bc} = (1/2)^3\Phi_{cc}$; for Figure 20(b),$\Phi_{abc} = (1/2)\Phi_{bbc} = (1/2)^2\Phi_{bc} = (1/2)^4\Phi_{AA}$; for Figure 20(c), $\Phi_{abc} = (1/2)^2\Phi_{ssc} = (1/2)^3\Phi_{sc} = (1/2)^5\Phi_{AA}$.

Using the path-counting formula (12), if a path-triple $\langle P_{Aa}, P_{Ab}, P_{Ac}\rangle$ has root overlap (i.e., Type 2), then the contribution of $\langle P_{Aa}, P_{Ab}, P_{Ac}\rangle$ to Φ_{abc} can be computed as follows:$\sum_{\text{Type 2}}(1/2)^{L_{\langle P_{Aa},P_{Ab},P_{Ac}\rangle}+1}\Phi_{AA}$, where $L_{\langle P_{Aa},P_{Ab},P_{Ac}\rangle} = L_{P_{Aa}} + L_{P_{Ab}} + L_{P_{Ac}} - L_{P_{As}}$ and s is the last individual of the root overlap path P_{As}.

For Figure 20(a), c is the only triple-common ancestor and we obtain $\Phi_{abc} = (1/2)^{L_{\langle P_{ca},P_{cb},P_{cc}\rangle}+1}\Phi_{cc} = (1/2)^{2+1}\Phi_{cc} = (1/2)^3\Phi_{cc}$. Similarly, for Figures 20(b) and 20(c), we obtain $\Phi_{abc} = (1/2)^4\Phi_{AA}$ and $\Phi_{abc} = (1/2)^5\Phi_{AA}$, respectively.

Induction Step. Let n denote the number of edges in G. Assume true for $n \le k$, where $k \ge 2$. Show that it is true for $= k + 1$.

For Figures 21(a), 21(b), and 21(c), among a, b, and c, let a be the individual who has the longest path and let p be a parent of a. Then, we cut the edge $p \to a$ from G and obtain a new graph G^* which satisfies the condition of induction hypothesis. For Figure 21(a), we use the path-counting formula for Φ_{fbc} in G^* : $\Phi_{fbc} = \sum_{\text{Type 2}}(1/2)^{L_{\langle P_{Af},P_{Ab},P_{Ac}\rangle}+1}\Phi_{AA}$.

In G, f is the only parent of a, according to the recursive formula (3), we have $\Phi_{abc} = (1/2)\Phi_{fbc}$. Then, we can plug-in the Φ_{fbc} and obtain

$$\begin{aligned}
\Phi_{abc} &= \frac{1}{2}\Phi_{fbc} \\
&= \frac{1}{2}\sum_{\text{Type 2}}\left(\frac{1}{2}\right)^{L_{\langle P_{Af},P_{Ab},P_{Ac}\rangle}+1}\Phi_{AA} \\
&= \sum_{\text{Type 2}}\left(\frac{1}{2}\right)^{L_{\langle P_{Af},P_{Ab},P_{Ac}\rangle}+1+1}\Phi_{AA} \\
&\because L_{\langle P_{Aa},P_{Ab},P_{Ac}\rangle} = L_{\langle P_{Af},P_{Ab},P_{Ac}\rangle} + 1 \\
&\therefore \Phi_{abc} = \sum_{\text{Type 2}}\left(\frac{1}{2}\right)^{L_{\langle P_{Af},P_{Ab},P_{Ac}\rangle}+1+1}\Phi_{AA} \\
&= \sum_{\text{Type 2}}\left(\frac{1}{2}\right)^{L_{\langle P_{Aa},P_{Ab},P_{Ac}\rangle}+1}\Phi_{AA}.
\end{aligned} \tag{B.3}$$

For Figures 21(b) and 21(c), we take the same steps as we calculate Φ_{abc} for Figure 21(a).

In summary, it is true for $n = k + 1$. □

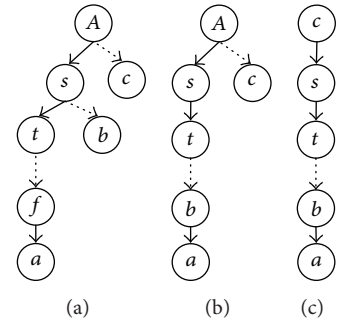


FIGURE 21: (a) No individual who is a parent of another; (b) b is a parent of a; (c) b is a parent of a and c is an ancestor of b.

B.1.3. Correctness Proof for Case 2.3

Case 2.3. For Φ_{aab}, the path-triples in the pedigree graph G have mergeable path-pair.

Proof. Considering the relationship between a and b, G has two scenarios: (i) b is not an ancestor of a; (ii) b is an ancestor of a. Using the path-counting formula (A.1), if a path-triple $\langle P_{Aa1}, P_{Aa2}, P_{Ab}\rangle \in$ Type 3, which means that it has a mergeable path-pair, then the contribution of $\langle P_{Aa1}, P_{Aa2}, P_{Ab}\rangle$ to Φ_{aab} can be computed as follows: $\sum_{\text{Type 3}}(1/2)^{L_{\langle P_{Aa},P_{Ab}\rangle}+1}\Phi_{AA}$, where $L_{\langle P_{Aa},P_{Ab}\rangle} = L_{P_{Aa}} + L_{P_{Ab}}$.

Using the recursive formula (4), we obtain $\Phi_{aab} = (1/2)(\Phi_{ab} + \Phi_{fmb})$.

For Figure 22(a), A is a common ancestor of a and b.

$\because a$ only has one parent f

$$\begin{aligned}
\therefore \Phi_{aab} &= \frac{1}{2}\left(\Phi_{ab} + \Phi_{fmb}\right) \\
&= \frac{1}{2}\left(\Phi_{ab} + 0\right) = \frac{1}{2}\Phi_{ab} \quad \left(\text{as } m \text{ is missing}\right).
\end{aligned} \tag{B.4}$$

For Φ_{ab}, we use Wright's formula and obtain $\Phi_{ab} = \sum_P(1/2)^{L_{\langle P_{Aa},P_{Ab}\rangle}}\Phi_{AA}$ where P denotes all nonoverlapping path-pairs $\langle P_{Aa}, P_{Ab}\rangle$.

Then, we have $\Phi_{aab} = (1/2)\Phi_{ab} = (1/2)\sum_P(1/2)^{L_{\langle P_{Aa},P_{Ab}\rangle}}\Phi_{AA} = \sum_P(1/2)^{L_{\langle P_{Aa},P_{Ab}\rangle}+1}\Phi_{AA}$.

For Figure 22(b), we can also transform the computation of Φ_{aab} to Φ_{ab}.

In summary, it shows that the path-counting formula (A.1) is true for Case 2.3. □

B.1.4. Correctness Proof for Cases 2.1 and 2.2. For Φ_{aab}, when there is no path-triple having mergeable path-pair, (i.e., the path-triple belongs to either Case 2.1 or Case 2.3), Φ_{aab} can be transformed to $\Phi_{a_1a_2b}$, which is equivalent to the computation of Φ_{abc} for Cases 3.1 and 3.2. The correctness of our path-counting formula for Cases 3.1 and 3.2 is proven. Thus, we obtain the correctness for Φ_{aab} when the path-triple belongs to either Case 2.1 or Case 2.2.

B.2. Multiple Triple-Common Ancestors. Now, we provide the correctness proof for multiple triple-common ancestors regarding the path-counting formulas (12) and (A.1).

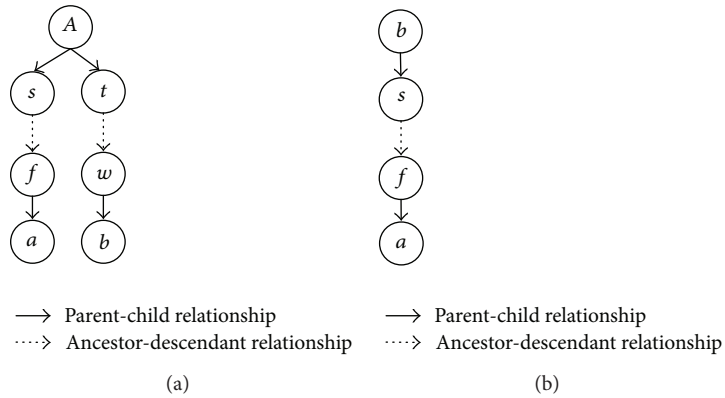


FIGURE 22: (a) b is not an ancestor of a; (b) b is an ancestor of a.

Lemma A. *Given a pedigree graph G and three individuals a, b, c having at least one trip-common ancestor, Φ_{abc} is correctly computed using the path counting formulas* (12) *and* (A.1).

Proof. Proof by induction on the number of triple-common ancestors

Basis. G has only one triple-common ancestor of a, b, and c.

The correctness of (12) and (A.1) for G with only one triple-common ancestor of a, b, and c is proven in the previous section.

Induction Hypothesis. Assume that if G has k or less triple-common ancestors of a, b, and c, (12) and (A.1) are correct for G.

Induction Step. Now, we show that it is true for G with $k+1$ triple-common ancestors of a, b, and c.

Let $Tri\_C(a,b,c,G)$ denote all triple-common ancestors of a, b, and c in G, where $Tri\_C(a,b,c,G) = \{A_i \mid 1 \le i \le k+1\}$. Let A_1 be the most top triple-common ancestor such that there is no individual among the remaining ancestors $\{A_i \mid 2 \le i \le k+1\}$ who is an ancestor of A_1. Let $S(A_1)$ denote the contribution from A_1 to Φ_{abc}.

Because A_1 is the most top triple-common ancestor, there is no path-triple from $\{A_i \mid 2 \le i \le k+1\}$ to a, b, and c which passes through A_1. Then, we can remove A_1 from G and delete all out-going edges from A_1 and obtain a new graph G' which has k triple-common ancestors of a, b, and c. It means $Tri\_C(a,b,c,G') = \{A_i \mid 2 \le i \le k+1\}$.

For the new graph G', we can apply our induction hypothesis and obtain $\Phi_{abc}(G')$.

For the most top triple-common ancestor A_1, there are two different cases considering its relationship with the other triple-common ancestors:

(1) there is no individual among $\{A_i \mid 2 \le i \le k+1\}$ who is a descendant of A_1;

(2) there is at least one individual among $\{A_i \mid 2 \le i \le k+1\}$ who is a descendant of A_1.

For (1), since no individual among $\{A_i \mid 2 \le i \le k+1\}$ is a descendant of A_1, the set of path-triples from A_1 to a, b, and c is independent of the set of path-triples from $\{A_i \mid 2 \le i \le k+1\}$ to a, b, and c. It also means that the contribution from A_1 to Φ_{abc} is independent of the contribution from the other triple-common ancestors.

Summing up all contributions, we can obtain $\Phi_{abc}(G) = \Phi_{abc}(G') + S(A_1)$.

For (2), let A_j be one descendant of A_1. Now both A_1 and A_j can reach a, b, and c.

$pt_i = \{t_a\colon A_1 \to \cdots \to a; t_b\colon A_1 \to \cdots \to b; t_c\colon A_1 \to \cdots \to c\}$, a path-triple from A_1 to a, b, and c.

If t_a, t_b, and t_c all pass through A_j, then the path-triple pt_i is not an eligible path-triple for Φ_{abc}. When we compute the contribution from A_1 to Φ_{abc}, we exclude all such path-triples where t_a, t_b, and t_c all pass through a lower triple-common ancestor. In other words, an eligible path-triple from A_1 regarding Φ_{abc} cannot have three paths all passing through a lower triple-common ancestor. Therefore, we know that that the contribution from A_1 to Φ_{abc} is independent of the contribution from the other triple-common ancestors. Summing up all contributions, we obtain $\Phi_{abc}(G) = \Phi_{abc}(G') + S(A_1)$. □

C. Proof for Four Individuals and Two Pairs of Individuals

Here, we give a proof sketch for the correctness of path counting formulas for four individuals. First of all, for four individuals in a pedigree graph G, we present all different cases based on which we construct a dependency graph. The correctness of the path-counting formulas for two-pair individuals can be proved similarly.

C.1. Proof for Four Individuals. Consider the existence of different types of path-quads regarding Φ_{abcd}, Φ_{aabc}, and Φ_{aaab}; there are 15 cases for a pedigree graph G:

$$\left.\begin{array}{l}\text{Case 2.1: } G \text{ has path-triples} \\ \quad \langle P_{Aa_1}, P_{Aa_2}, P_{Ab}\rangle \\ \quad \text{with zero root overlap} \\ \text{Case 2.2: } G \text{ has path-triples} \\ \quad \langle P_{Aa_1}, P_{Aa_2}, P_{Ab}\rangle \\ \quad \text{with one root overlap} \\ \text{Case 2.3: } G \text{ has path-pairs} \\ \quad \langle P_{Aa}, P_{Ab}\rangle \\ \quad \text{with zero root overlap}\end{array}\right\} \Longleftarrow \Phi_{aaab},$$

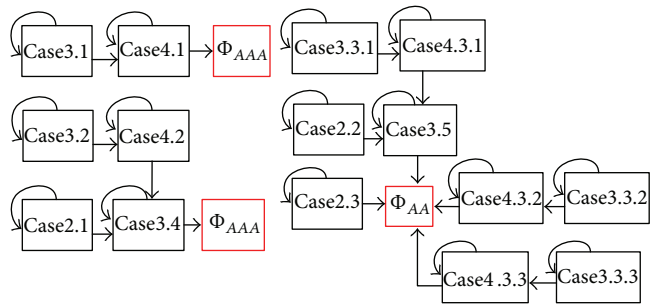


Figure 23: Dependency graph for different cases for four individuals.

$$\left.\begin{array}{l}\text{Case 3.1: } G \text{ has path-quads } \langle P_{Aa_1}, P_{Aa_2}, P_{Ab}, P_{Ac}\rangle \text{ with zero root overlap}\\ \text{Case 3.2: } G \text{ has path-quads } \langle P_{Aa_1}, P_{Aa_2}, P_{Ab}, P_{Ac}\rangle \text{ with one root 2-overlap}\\ \text{Case 3.3.1: } G \text{ has path-quads } \langle P_{Aa_1}, P_{Aa_2}, P_{Ab}, P_{Ac}\rangle \text{ with two root 2-overlap}\\ \text{Case 3.3.2: } G \text{ has path-quads } \langle P_{Aa_1}, P_{Aa_2}, P_{Ab}, P_{Ac}\rangle \text{ with one root 3-overlap}\\ \text{Case 3.3.3: } G \text{ has path-quads } \langle P_{Aa_1}, P_{Aa_2}, P_{Ab}, P_{Ac}\rangle \text{ with one root 2-overlap and one root 3-overlap}\\ \text{Case 3.4: } G \text{ has path-triples } \langle P_{Aa}, P_{Ab}, P_{Ac}\rangle \text{ with zero root overlap}\\ \text{Case 3.5: } G \text{ has path-triples } \langle P_{Aa}, P_{Ab}, P_{Ac}\rangle \text{ with one root overlap}\end{array}\right\} \Longleftarrow \Phi_{aabc},$$

$$\left.\begin{array}{l}\text{Case 4.1: } G \text{ has path-quads } \langle P_{Aa}, P_{Ab}, P_{Ac}, P_{Ad}\rangle \text{ with zero root overlap}\\ \text{Case 4.2: } G \text{ has path-quads } \langle P_{Aa}, P_{Ab}, P_{Ac}, P_{Ad}\rangle \text{ with one root 2-overlap}\\ \text{Case 4.3.1: } G \text{ has path-quads } \langle P_{Aa}, P_{Ab}, P_{Ac}, P_{Ad}\rangle \text{ with two root 2-overlap}\\ \text{Case 4.3.2: } G \text{ has path-quads } \langle P_{Aa}, P_{Ab}, P_{Ac}, P_{Ad}\rangle \text{ with one root 3-overlap}\\ \text{Case 4.3.3: } G \text{ has path-quads } \langle P_{Aa}, P_{Ab}, P_{Ac}, P_{Ad}\rangle \text{ with one root 2-overlap and one root 3-overlap}\end{array}\right\} \Longleftarrow \Phi_{abcd}. \quad \text{(C.1)}$$

Then, we construct a dependency graph shown in Figure 23 for all cases for four individuals.

According to the dependency graph in Figure 23, the intermediate steps including Cases 3.4 and 3.5 are already proved for the computation of Φ_{abc}. The correctness of the transformation from Case 4.2 to Case 3.4 can be proved based on the recursive formula for Φ_{abcd} and Φ_{aabc}. Similarly, we can obtain the transformation from Case 4.3.1 to Case 3.5.

C.2. Proof for Two Pairs of Individuals. Consider the existence of different types of 2-pair-path-pair regarding $\Phi_{ab,cd}$; there are 9 cases which are listed as follows.

Case 4.1. G has $\langle (P_{Aa}, P_{Ab}), (P_{Ac}, P_{Ad})\rangle$ with zero root homo-overlap and zero root heter-overlap.

Case 4.2. G has $\langle (P_{Aa}, P_{Ab}), (P_{Ac}, P_{Ad})\rangle$ with zero root homo-overlap and one root heter-overlap.

Case 4.3.1. G has $\langle (P_{Aa}, P_{Ab}), (P_{Ac}, P_{Ad})\rangle$ with zero root homo-overlap and two root heter-overlap.

Case 4.3.2. G has $\langle (P_{Aa}, P_{Ab}), (P_{Ac}, P_{Ad})\rangle$ with one root homo-overlap and two root heter-overlap.

Case 4.4. G has $\langle (P_{Aa}, P_{Ab}), (P_{Ac}, P_{Ad})\rangle$ with one root homo-overlap and zero root heter-overlap.

Case 4.5. G has $\langle (P_{Aa}, P_{Ab}), (P_{Ac}, P_{Ad})\rangle$ with two root homo-overlap and zero root heter-overlap.

Case 4.6. G has path-triples $\langle P_{Aa}, P_{Ab}, P_{Ac}\rangle$ with zero root overlap.

Case 4.7. G has path-triples $\langle P_{Aa}, P_{Ab}, P_{Ac}\rangle$ with one root overlap.

Case 4.8. G has path-pairs $\langle P_{Tc}, P_{Td}\rangle$ with zero root overlap.

Then, we construct a dependency graph for the cases relating to $\Phi_{ab,cd}$ in Figure 24.

According to the dependency graph in Figure 24, Cases 4.6, 4.7, and 4.8 are the intermediate steps which already are proved for the computation of Φ_{abc}. The correctness of the transformation from Case 4.2 to Case 4.6 can be proved based on the recursive formula for $\Phi_{ab,cd}$ and $\Phi_{ab,ac}$. Similarly, we can obtain the transformation from Cases 4.3.1 and 4.3.2 to Case 4.7 as well as from Case 4.4 to Case 4.8 accordingly.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

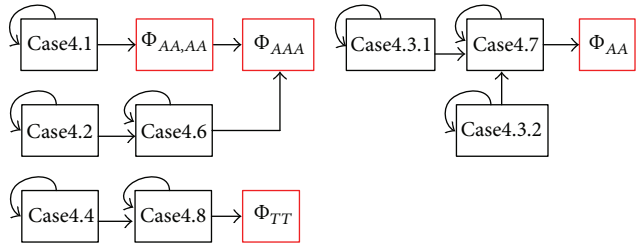


Figure 24: Dependency graph for different cases for two pairs of individuals.

Acknowledgments

The authors thank Professor Robert C. Elston, Case School of Medicine, for introducing to them the identity coefficients and referring them to the related literature [7, 10, 17]. This work is partially supported by the National Science Foundation Grants DBI 0743705, DBI 0849956, and CRI 0551603 and by the National Institute of Health Grant GM088823.

References

[1] Surgeon General's New Family Health History Tool Is Released, Ready for "21st Century Medicine", http://compmed.com/category/people-helping-people/page/7/.

[2] M. Falchi, P. Forabosco, E. Mocci et al., "A genomewide search using an original pairwise sampling approach for large genealogies identifies a new locus for total and low-density lipoprotein cholesterol in two genetically differentiated isolates of Sardinia," *The American Journal of Human Genetics*, vol. 75, no. 6, pp. 1015–1031, 2004.

[3] M. Ciullo, C. Bellenguez, V. Colonna et al., "New susceptibility locus for hypertension on chromosome 8q by efficient pedigree-breaking in an Italian isolate," *Human Molecular Genetics*, vol. 15, no. 10, pp. 1735–1743, 2006.

[4] Glossary of Genetic Terms, National Human Genome Research Institute, http://www.genome.gov/glossary/?id=148.

[5] C. W. Cotterman, *A calculus for statistico-genetics [Ph.D. thesis]*, Columbus, Ohio, USA, Ohio State University, 1940, Reprinted in P. Ballonoff, Ed., *Genetics and Social Structure*, Dowden, Hutchinson & Ross, Stroudsburg, Pa, USA, 1974.

[6] G. Malecot, *Les mathématique de l'hérédité*, Masson, Paris, France, 1948, Translated edition: *The Mathematics of Heredity*, Freeman, San Francisco, Calif, USA, 1969.

[7] M. Gillois, "La relation d'identité en génétique," *Annales de l'Institut Henri Poincaré B*, vol. 2, pp. 1–94, 1964.

[8] D. L. Harris, "Genotypic covariances between inbred relatives," *Genetics*, vol. 50, pp. 1319–1348, 1964.

[9] A. Jacquard, "Logique du calcul des coefficients d'identite entre deux individuals," *Population*, vol. 21, pp. 751–776, 1966.

[10] G. Karigl, "A recursive algorithm for the calculation of identity coefficients," *Annals of Human Genetics*, vol. 45, no. 3, pp. 299–305, 1981.

[11] B. Elliott, S. F. Akgul, S. Mayes, and Z. M. Ozsoyoglu, "Efficient evaluation of inbreeding queries on pedigree data," in *Proceedings of the 19th International Conference on Scientific and Statistical Database Management (SSDBM '07)*, July 2007.

[12] B. Elliott, E. Cheng, S. Mayes, and Z. M. Ozsoyoglu, "Efficiently calculating inbreeding on large pedigrees databases," *Information Systems*, vol. 34, no. 6, pp. 469–492, 2009.

[13] L. Yang, E. Cheng, and Z. M. Özsoyoğlu, "Using compact encodings for path-based computations on pedigree graphs," in *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine (ACM-BCB '11)*, pp. 235–244, August 2011.

[14] E. Cheng, B. Elliott, and Z. M. Ozsoyoglu, "Scalable computation of kinship and identity coefficients on large pedigrees," in *Proceedings of the 7th Annual International Conference on Computational Systems Bioinformatics (CSB '08)*, pp. 27–36, 2008.

[15] E. Cheng, B. Elliott, and Z. M. Özsoyoğlu, "Efficient computation of kinship and identity coefficients on large pedigrees," *Journal of Bioinformatics and Computational Biology (JBCB)*, vol. 7, no. 3, pp. 429–453, 2009.

[16] S. Wright, "Coefficients of inbreeding and relationship," *The American Naturalist*, vol. 56, no. 645, 1922.

[17] R. Nadot and G. Vaysseix, "Kinship and identity algorithm of coefficients of identity," *Biometrics*, vol. 29, no. 2, pp. 347–359, 1973.

[18] E. Cheng, *Scalable path-based computations on pedigree data [Ph.D. thesis]*, Case Western Reserve University, Cleveland, Ohio, USA, 2012.

[19] V. Ollikainen, *Simulation Techniques for Disease Gene Localization in Isolated Populations [Ph.D. thesis]*, University of Helsinki, Helsinki, Finland, 2002.

[20] H. T. T. Toivonen, P. Onkamo, K. Vasko et al., "Data mining applied to linkage diseqilibrium mapping," *The American Journal of Human Genetics*, vol. 67, no. 1, pp. 133–145, 2000.

[21] W. Boucher, "Calculation of the inbreeding coefficient," *Journal of Mathematical Biology*, vol. 26, no. 1, pp. 57–64, 1988.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 956917, 4 pages
http://dx.doi.org/10.1155/2014/956917

Research Article

A Note regarding Problems with Interaction and Varying Block Sizes in a Comparison of Endotracheal Tubes

Richard L. Einsporn[1] and Zhenyu Jia[1,2]

[1] *Department of Statistics, The University of Akron, 302 Buchtel Common, Akron, OH 44325-1913, USA*
[2] *Department of Family and Community Medicine, Northeast Ohio Medical University, Rootstown, OH 44272, USA*

Correspondence should be addressed to Richard L. Einsporn; rle@uakron.edu

Received 31 January 2014; Accepted 1 July 2014; Published 15 July 2014

Academic Editor: Xiao-Qin Xia

A randomized clinical experiment to compare two types of endotracheal tubes utilized a block design where each of the six participating anesthesiologists performed tube insertions for an equal number of patients for each type of tube. Five anesthesiologists intubated at least three patients with each tube type, but one anesthesiologist intubated only one patient per tube type. Overall, one type of tube outperformed the other on all three effectiveness measures. However, analysis of the data using an interaction model gave conflicting and misleading results, making the tube with the better performance appear to perform worse. This surprising result was caused by the undue influence of the data for the anesthesiologist who intubated only two patients. We therefore urge caution in interpreting results from interaction models with designs containing small blocks.

1. Introduction

A clinical research investigation by Radesic et al. [1] compared two types of endotracheal tubes (ETTs) used by anesthesiologists. The original plan for the study utilized a generalized randomized block design [2, 3] (stratified allocation), in which each of six anesthesiology providers (hereafter "APs") was to use one type of tube for five patients and the other type of tubes for five patients, with assignment of patient to tube being randomized. Three dependent variables obtained for each patient were used to compare the types of tubes: time to complete the intubation, number of times the insertion had to be momentarily stopped and the tube redirected, and a rating by the AP of the difficulty of the insertion. It was anticipated that there could be some interactive effect between the type of tube and the AP with respect to these response variables, in that the differences between the tube types could vary according to the APs' proficiencies and preferences.

In the course of conducting this study, it turned out that some of the APs who were enlisted to participate were seldom available, while others were frequently available. In order to complete the investigation within an allotted time frame, the number of patients per AP was altered with more than ten patients for some APs and fewer than ten for others. Still, each AP had an even number of patients with half being randomized to each type of tube. One particular AP had only two patients, one per tube type. In the original analysis presented in Radesic et al. [1], the researchers deemed this AP to have done too few intubations and excluded that data from the analysis. A further analysis that did include the data for this AP revealed a spurious result that conflicts with the conclusions of the original study. It is this contradictory finding that is the focus of this paper. Such a result should sound a note of caution to data analysts who include interaction terms in their models.

In this paper, we first provide some additional details of both the design and original analysis of the anesthesiology tube study by Radesic et al. Then we will illustrate the specific problem that arises when an interaction term is added to the statistical model. Finally, we discuss how such a problem could arise in many other situations where an interaction term may be included in a model.

2. Materials and Methods

The purpose of the study by Radesic et al. [1] was to compare the performance of two types of ETTs when used in conjunction with the GlideScope, a video laryngoscope.

Table 1: Intubation outcomes for the Parker Flex-Tip and standard tubes.

Dependent variables	Parker Flex-Tip ($n = 30$) mean (SD)	Standard ETT ($n = 30$) mean (SD)
Time for ETT insertion (sec)	10.9 (7.5)	12.4 (7.3)
Number of redirections	0.7 (1.5)	1.3 (2.7)
Difficulty of insertion rating	14.3 (14.9)	17.4 (19.7)

The Parker Flex-Tip (PFT) and the standard Mallinckrodt were the two types of ETTs used in this study. The GlideScope allows the AP to visualize the airway structures when passing the ETT into the oral pharynx, through the glottis, and into the trachea.

Six APs and 60 patients participated in the study. In the modified design, one AP intubated 22 patients, another intubated 18 patients, three APs performed 6 intubations each, and one AP only performed two intubations. The APs were balanced with respect to the ETT type, in that half of each AP's intubations were done with the PFT tube and half with the standard Mallinckrodt tube. In the original analysis [1], the data for the AP who did only two intubations was discarded, leaving a sample size of 58 patients, utilizing data for only five APs. The three dependent variables were (1) time for ETT insertion, (2) number of ETT redirections, and (3) ease of use rating by the AP immediately following each intubation. Values for the first two dependent variables were determined precisely by means of viewing a video recording of each intubation. To rate the ease of use, a 100 mm visual analog scale (VAS) was used, with 0 representing "easiest insertion" and 100 representing "hardest insertion." After each intubation, the AP made a mark along this 100mm line to rate the difficulty of the insertion.

The analysis presented in [1] utilized data for only the 58 patients who were intubated by the 5 APs who did six or more intubations, excluding the AP who did only two intubations. A two-factor ANCOVA model was used, with ETT type and AP being the two designed factors and two patient characteristic variables serving as covariates. These were the Cormack-Lehane view (2 categories) and whether the muscles were paralyzed, as determined by observation of nerve stimulation. The model included interaction terms for the ETT type with each covariate and with the AP factor. The AP was entered into the model as a random effect. Two of the dependent variables were transformed using logs in order to correct for skewness.

When the results were averaged for the 58 patients (aggregated over the five APs and the covariates), the PFT tube had lower (better) mean responses on each of the dependent variables. Likewise, for all three dependent variables, the adjusted means resulting from the model described above were lower for the PFT. P values for two of the dependent variables—time to intubate and difficulty rating—were below .01.

In this paper, we will do a similar analysis, this time using the data for all 60 patients and all six APs. To make our point in the most straightforward fashion, our analysis will exclude the two covariates. For the same reason, we will keep the dependent variables in their original units, rather than using log transformations. (The presence of covariates in the model or the use of transformed data does not change the essence of the results.)

Table 2: Mean intubation outcomes for the Parker Flex-Tip and standard tubes for each of the six anesthesiology providers.

Dependent variables	Parker Flex-Tip	Standard ETT
AP#1	$N = 3$	$N = 3$
Time for ETT insertion (sec)	9.0	14.0
Number of redirections	1.0	2.0
Difficulty of insertion rating	16.7	19.0
AP#2	$N = 11$	$N = 11$
Time for ETT insertion (sec)	6.7	9.9
Number of redirections	0.0	0.7
Difficulty of insertion rating	3.7	11.5
AP#3	$N = 9$	$N = 9$
Time for ETT insertion (sec)	14.7	17.1
Number of redirections	1.6	2.4
Difficulty of insertion rating	21.7	31.8
AP#4	$N = 1$	$N = 1$
Time for ETT insertion (sec)	15.0	5.0
Number of redirections	3.0	0.0
Difficulty of insertion rating	60.0	7.0
AP#5	$N = 3$	$N = 3$
Time for ETT insertion (sec)	8.0	6.0
Number of redirections	1.0	0.0
Difficulty of insertion rating	13.3	11.7
AP#6	$N = 3$	$N = 3$
Time for ETT insertion (sec)	18.7	14.7
Number of redirections	0.0	0.7
Difficulty of insertion rating	14.2	3.0

3. Results and Discussion

Table 1 shows the mean values for each of the dependent variables when the data are aggregated over all six APs. For each dependent variable, the mean response is lower (better) for the PFT tube than for the standard ETT. The results are presented separately for each AP in Table 2. It can be seen that the fourth anesthesiology provider (AP #4) had one patient who was difficult to intubate and one for whom intubation was very easy. Whether this is due to the type of tube or to patient characteristics cannot be sorted out statistically due to confounding.

Table 3: Least squares adjusted means for each type of tube using an additive model or an interaction model.

Dependent variables	Additive model		Interaction model	
	Parker Flex-Tip	Standard ETT	Parker Flex-Tip	Standard ETT
Time for ETT insertion (sec)	10.8	12.3	12.0	11.1
Number of redirections	0.8	2.3	1.1	1.0
Difficulty of insertion rating	16.3	19.4	21.6	14.0

Note. The adjusted means were obtained using the General Linear Model ANOVA platform in Minitab V.16 and are the same as those obtained using PROC GLM in SAS V.9.3.

3.1. Additive Model versus Interactive Model Results. First, consider the results of an additive model in which the factors are tube type (fixed) and anesthesiology provider (random). Such a model will allow us to compare the tube types, while adjusting for potential differences among the APs with respect to the dependent variables. For example, some APs could be faster at performing intubations than others. Variation in the dependent variables due to AP differences would then be accounted for and removed from the "error term" used for comparing the tube types. Univariate two-way ANOVAs were run for each of the three dependent variables. According to ANOVA F-tests, the difference between the PFT tube and standard tube was not found to be statistically significant for any of the three effectiveness measures. (This is also true if the data for AP #4 are removed.) However, the adjusted mean for the PFT tube was lower (better) than for the standard tube on each of the three dependent variables (Table 3).

In order to allow for the possibility that the differences between the tube types may vary among APs, an interaction term was added to the model. For example, some APs may tend to perform better with one tube while other APs do better with the other tube. Again, univariate ANOVAs were run for each of the three dependent variables, this time with the interaction term, tube type $*$ AP, included in the model. Surprisingly, the adjusted means resulting from these analyses make it appear that the PFT tube performs worse than the standard tube (Table 3). Again, differences are not statistically significant according to the ANOVA F-tests.

The adjusted means shown in Table 3 were produced using the General Linear Model ANOVA platform in Minitab V.16 and are the same as those obtained using PROC GLM in SAS V.9.3. To its credit, Minitab's default output flags both of the data points for AP #4 as having "large leverage" for both the additive and interaction models. We also note that the same adjusted means are produced even if the APs are entered into the model as fixed rather than random effects.

We believe that the results obtained using the interaction model are misleading due to the undue influence of the results for the one AP who intubated only one patient with each type of ETT. Further, we were somewhat surprised by this, because the design was balanced in the sense that each AP used each ETT type the same number of times, meaning that the ETT and AP factors are orthogonal in the design matrix.

3.2. Illustration. The misleading results obtained in the ETT study could arise in many similar situations. Here is a simple example to illustrate the problem in the context of a two-factor factorial analysis. Suppose that Factors A and B have a and b levels, respectively, and that, within each level of Factor A, the same number of observations is obtained for each level of B, although this number may vary among the levels of A. As in the anesthesia tube study, we are primarily focusing on the impact of only one factor, here, Factor B.

Table 4: Hypothetical data for a two-factor study.

	Factor B level 1	Factor B level 2
Factor A level 1	10 11 12 11 10	8 6 5 7 7
Factor A level 2	9 11 12 10	6 7 4 5
Factor A level 3	4	15

Table 5: Raw and adjusted means for the two levels of B for the hypothetical data.

	Factor B level 1	Factor B level 2
Raw means	10.00	7.00
Adjusted means; additive model	10.23	7.23
Adjusted means; interaction model	8.43	9.03

If n_{ij} represents the number of observations for the ith level of A and jth level of B, then $n_{i1} = n_{i2} = \cdots = n_{ib}$ for $i = 1, 2, \ldots, a$. Consider the case where $a = 3$ and $b = 2$; $n_{1j} = 5$, $n_{2j} = 4$, and $n_{3j} = 1$, $j = 1, 2$. Suppose the values of the dependent variable are as shown in Table 4.

In this case, the raw means for the two levels of Factor B differ by 3.0 with the B1 mean higher than the B2 mean (Table 5). Using an additive model, PROC GLM in SAS produces adjusted means that are also 3.0 units apart, and the difference is statistically significant ($P = .030$). For the interaction model, the adjusted means for the two levels of B are reversed in order of magnitude, though the difference is not statistically significant ($P = .367$ using SAS type III sums of squares).

Minitab's General Linear Model ANOVA produces the same results for both the additive and interaction models. To its credit, Minitab also issues a warning in its output that the two observations for A3 have high leverage. To investigate this further, we performed regression analyses, which allowed us to assess the leverage and influence of the two data values for A3. To do this, we created indicator variables for A1, A2, and B1 and multiplicative interaction terms A1 $*$ B1 and A2 $*$ B1. Then we ran a regression analysis with both an additive model

(Y versus A1, A2, and B1) and an interaction model (Y versus A1, A2, B1, A1 ∗ B1, and A2 ∗ B1), requesting that influence diagnostics be included in the output (INFLUENCE option in PROC REG). For the additive model, each of the two A3 observations had somewhat high leverage (hat diagonal = .55) and strong influence on the estimate for B1 (DFBETAS = −3.11) (see Belsley et al. [4]). However, for the interaction model, these two points had the maximum possible leverage (hat diagonals = 1.00) and extreme influence on all the coefficient estimates (DFBETAS all infinite/undefined). With only one observation at each combination of A3 and B, it is clear that an interactive model will fit the response variable exactly and thus the maximum leverage.

4. Conclusion

There are many clinical studies, such as the ETT comparison described here, where allocation of patients to treatments may be blocked or stratified (see [5] for a discussion of stratification in the clinical trial setting). For example, patients may be stratified by center, race, or disease status. In such cases, additive models for comparing the treatments will properly adjust for net differences in the dependent variables for the different strata. However, it may make sense for an interaction model to be used in the model as well. For example, the benefit afforded by one treatment over another may be greater for one racial group than for another. If one or more of the strata are very small in size, then the phenomenon illustrated by the examples of this paper suggests caution be used in interpretation of the results. Data for the small strata may have undue influence on the findings, since these observations will have high leverage. As shown here, this problem holds even for the "unbiased" case where, within any stratum, an equal number of subjects receive each treatment. In light of these observations, we recommend that strata or blocks of size two be omitted from the data if an interaction model is used. This advice was followed in the original ETT comparison analysis [1], where the AP who intubated only two patients was removed from the data.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] B. P. Radesic, C. Winkelman, R. Einsporn, and J. Kless, "Ease of intubation with the Parker Flex-Tip or a standard Mallinckrodt endotracheal tube using a video laryngoscope (GlideScope)," *AANA Journal*, vol. 80, no. 5, pp. 363–372, 2012.

[2] S. Addelman, "The generalized randomized block design," *The American Statistician*, vol. 23, no. 4, pp. 35–36, 1969.

[3] K. Hinkelmann and O. Kempthorne, *Design and Analysis of Experiments, Volume I: Introduction to Experimental Design*, John Wiley & Sons, New York, NY, USA, 2nd edition, 2008.

[4] D. A. Belsley, E. Kuh, and R. E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley, 2004.

[5] D. Harrington, *Designs for Clinical Trials: Perspectives on Current Issues*, Springer, New York, NY, USA, 2012.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 758718, 9 pages
http://dx.doi.org/10.1155/2014/758718

Research Article

A Mixture Modeling Framework for Differential Analysis of High-Throughput Data

Cenny Taslim and Shili Lin

Department of Statistics, The Ohio State University, Columbus, OH 43210, USA

Correspondence should be addressed to Shili Lin; shili@stat.osu.edu

Received 27 January 2014; Accepted 15 May 2014; Published 25 June 2014

Academic Editor: Samsiddhi Bhattacharjee

The inventions of microarray and next generation sequencing technologies have revolutionized research in genomics; platforms have led to massive amount of data in gene expression, methylation, and protein-DNA interactions. A common theme among a number of biological problems using high-throughput technologies is differential analysis. Despite the common theme, different data types have their own unique features, creating a "moving target" scenario. As such, methods specifically designed for one data type may not lead to satisfactory results when applied to another data type. To meet this challenge so that not only currently existing data types but also data from future problems, platforms, or experiments can be analyzed, we propose a mixture modeling framework that is flexible enough to automatically adapt to any moving target. More specifically, the approach considers several classes of mixture models and essentially provides a model-based procedure whose model is adaptive to the particular data being analyzed. We demonstrate the utility of the methodology by applying it to three types of real data: gene expression, methylation, and ChIP-seq. We also carried out simulations to gauge the performance and showed that the approach can be more efficient than any individual model without inflating type I error.

1. Introduction

With the completion of the human genome project more than a decade ago, large-scale approaches to biological research are advancing rapidly. In particular, the inventions of microarray and next generation sequencing technologies have revolutionized research in genomics; such high-throughput platforms have led to massive amount of data. Depending on the study, each type of experiment generates data with different characteristics. Among them are cDNA microarrays or RNA-seq for measuring changes in expression levels of thousands of genes simultaneously [1, 2]; ChIP-chip tiling arrays or ChIP-seq for studying genome-wide protein-DNA interactions [3, 4]; and differential methylation hybridization microarrays or whole genome bisulfite sequencing for performing whole genome DNA methylation profiling study [5, 6]. A common theme of interest for biologists when they employ these experiments is to perform differential analysis [7–12]. For example, in gene expression profiling, be it microarray or sequencing based, there is an interest in finding genes that are differentially expressed. For epigenetic profiling of cancer samples, it is of interest to find CpG islands that are differentially methylated between cancerous and normal cells. On the other hand, ChIP-seq data are frequently used to interrogate protein binding differentiation under two different conditions. Over the past decade, methods have been proposed for each type of data when new platforms/technologies were launched. Despite the common theme, different data types have their own unique features, creating a "moving target" scenario. As such, methods specifically designed for one data type may not lead to satisfactory results when applied to another data type. Furthermore, new data types from new biological experiments will continue to emerge as we are entering a new era of discovery [13, 14]. As such, it would be desirable to have a unified approach that would provide satisfactory solutions to multitype data, both those currently available and those that will become available in the future. To meet this challenge so that not only currently existing data types but also data from future problems, platforms, or experiments can be analyzed, we propose a mixture modeling framework that is flexible enough to automatically adapt to any moving

target. That is, the model we are proposing is adaptive to the data being analyzed rather than being fixed. More specifically, the approach considers several classes of mixture models and essentially provides a model-based procedure with the following features: (1) use of an ensemble of multiclass models, (2) models within each class adapting to the data being analyzed, and (3) flexible scheme for component classification. Thus, depending on the underlying distribution of the data being analyzed, the model will adapt accordingly to provide the best fit, which, as we demonstrate through simulation, can lead to improved power and sensitivity of differential identification without inflating type I error. To illustrate the utility of the method, we employ it to analyze three diverse types of high-throughput data, each of which has led to improved fit compared to a single-model analysis.

2. Materials and Methods

2.1. Synopsis of the Ensemble Approach. Mixture model-based approaches have been proposed specifically for different data types. Here, we propose an approach that tries to synthesize the advantages of these approaches into one single package. Depending on the data being analyzed, this ensemble approach will select the model that best fits the data and perform model-based classification. The first mixture model being considered for the ensemble is the gamma-normal-gamma (GNG) model proposed for analyzing DNA methylation data [15]. It uses a special case of the gamma distribution (exponential) to capture data coming from differential group and utilizes multiple normal components to capture the nondifferentiating methylated group allowing for small biases even after normalization. We integrate this model with uniform-normal mixture model (NUDGE) proposed by Dean and Raftery [16] which uses one uniform and one normal component to analyze gene expression data. To extend the applicability of the ensemble approach to other omic data types, we add to this ensemble an extension of NUDGE, which we call iNUDGE, to improve the fit by following the idea from GNG using multiple normal components. A robust weighting scheme for GNG was also extended to (i)NUDGE. In addition, we allow some of the normal components to be classified as capturing differentiated observations based on their locations and scale parameters, further increasing the flexibility of the ensemble model. We note that this feature differs from the intended use of the normal component(s) in GNG and NUDGE. Depending on the underlying distribution of the data, the best overall model among the three classes will be selected and used for inferences. The ensemble nature of this procedure makes it highly adaptable to data from various platforms. We demonstrate this capability by applying it to three types of data: gene expression, DNA methylation, and ChIP-seq. In what follows, we describe our ensemble model, parameters estimation, model selection, and finally model-based classification.

2.2. Ensemble of Finite Mixture Models. In the proposed ensemble approach, we integrate advantages from different models by considering multiple underlying distributions. Specifically, a collection of three classes of mixture models are utilized. Each class of models is designed to fit the normalized data that are usually expressed as (log) differences under two experimental conditions, for example, healthy versus diseased or before versus after treatment.

Let $f(y)$ be the unknown density function of the normalized data point y, which is modeled as

$$f\left(y;\Psi\right) = \left(1-\pi\right) f_0\left(y;\Psi_0\right) + \pi f_1\left(y;\Psi_1\right), \tag{1}$$

where Ψ, Ψ_0, and Ψ_1 are the underlying model parameters for the mixture and each of the two components, respectively, and will be specified as the formulation enfolds. In this first level of mixture, f_1 is designated to capture differential elements (overdispersion) whereas f_0 is for those that are more centrally located. Nevertheless, f_0 may also be used to identify differential observations, as detailed in the second level of mixture modeling. Specifically, we model f_1 and f_0 as follows:

$$f_1\left(y;\Psi_1\right) = \begin{cases} U_{[a,b]}\left(y\right) & \text{for (i)NUDGE} \\ \rho E_1\left(-y \times I\left\{y < -\xi_1\right\};\beta_1\right) & \\ \quad + \left(1-\rho\right) & \\ \quad \times E_2\left(y \times I\left\{y > \xi_2\right\};\beta_2\right) & \text{for GNG,} \end{cases}$$

$$f_0\left(y;\Psi_0\right) = \begin{cases} N\left(y;\mu,\sigma^2\right) & \text{for NUDGE} \\ \sum_{k=1}^{K}\gamma_k & \\ \quad \times N\left(y;\mu_k,\sigma_k^2\right), & \\ \sum_{k=1}^{K}\gamma_k = 1 & \text{for iNUDGE and GNG.} \end{cases} \tag{2}$$

As we can see from the above modeling, the overdispersion in the data is captured by either a uniform distribution or a mixture of two exponential distributions (special case of gamma). The parameters of the uniform distributions, a and b, are part of the model parameters (i.e., $a, b \in \Psi_1$) and so are the scale parameters and the mixing proportion of the exponential distributions (i.e., $\rho, \beta_1, \beta_2 \in \Psi_1$). The location parameters, ξ_1 and ξ_2 both >0, are assumed to be known. In practice, $\widehat{\xi}_1 = |\max(y < 0)|$ and $\widehat{\xi}_2 = |\min(y > 0)|$ may be used as estimates of ξ_1 and ξ_2. The more centrally located data are represented by either a single normal distribution or a mixture of normal distributions. The location and scale parameters are part of the model parameters, that is, $\mu, \sigma^2, \mu_k, \sigma_k^2 \in \Psi_0$, and so are the mixing proportions γ_k and the number of components in the mixture, K; that is $\gamma_k, K \in \Psi_0$. Thus, $\Psi = \{\pi\} \cup \Psi_0 \cup \Psi_1$. Finally, $I\{\cdot\}$ is the usual indicator function that is equal to 1 if the condition in $\{\}$ is satisfied; otherwise, it is 0. Since any distribution can be well represented by a mixture of normal distributions, both f_1 and some components of the normal mixture will be designated as "differential" components, as detailed below.

2.3. Robust Parameter Estimation. In order to get a robust estimation of model parameters, following GNG, we use a weighted likelihood function in our ensemble model:

$$l(\Psi) = \sum_{i=1}^{n} w_i \log f(y_i; \Psi), \tag{3}$$

where y_i, for $i = 1, 2, \ldots, n$, are the normalized observed data and $0 \le w_i \le 1$ are some prespecified weights.

Weighted likelihood is used because we want to downgrade the contributions from observations with small "intensities." For example, in modeling log-ratio, we want to distinguish data points with the same log-ratio but vastly different magnitudes in their individual intensities. If we let u be the average log intensities (standardized to mean zero and standard deviation 1), then the lower half Huber's weight function:

$$w(u) = \begin{cases} 1, & \text{if } u > -c \\ \dfrac{c}{|u|} & \text{if } u \le -c, \end{cases} \tag{4}$$

where $c = 1.345$, can be used to downweigh those with smaller average intensities. In addition to Huber's weight function, Tukey's bisquare function may also be used [17]. Further, an upper half or a two-tailed weight function can be used if justifiable for a particular data type or study goal.

The EM algorithm is used to fit each class of models under the ensemble. The stopping criteria for our EM algorithm are when either $\|\Psi_{(m+1)} - \Psi_{(m)}\| < \epsilon$ or a maximum number of iterations M are reached. In our simulation and data analysis, we set $\epsilon = 10^{-5}$ and $M = 2000$, which are also the default setting in the program implementing the ensemble approach.

2.4. Model Selection and Model-Based Classification. In both GNG and iNUDGE models, we first need to determine K, the number of normal components in the model, also known as the order of the model. In our analysis, we examine models with $K = 1, 2, \ldots$ and choose K that maximizes the Bayesian information criterion (BIC [18]). We use BIC as it is in favor of parsimonious model since the penalty for additional parameters is stronger than the Akaike information criterion (AIC [19]). That is, when selecting the order of the model, we want to be extra careful not to choose models that are too complex. After identifying the best model within each class, we use the AIC to select the overall best model among the three classes. The use of this balanced model selection approach is not only to prevent the selection of a model that is too complex (thus using BIC within each class) but, in the meantime, also to avoid choosing a model that is overly simple (thus using AIC when selecting among the classes).

Using the best model selected, a two-step approach is taken to classify each observation as differential or not. In the first step, we classify a normal component $N(\mu_k, \sigma_k^2)$ as a differential one if one of its tails captures observations that are "outliers" in the overall distribution:

$$|\mu_k| + 2 \times \sigma_k > 1.5 \times \text{IQR}, \tag{5}$$

where IQR is the interquartile range of the entire dataset. The normal components that are not labeled as differential are called "nondifferential."

After each normal component is labeled, we compute the local false discovery rate (fdr) proposed by Efron [20] and adapted by Khalili et al. [15] for each observation:

$$\text{fdr}(y_i) = \frac{f_{\text{nd}}(y_i, \widehat{\Psi}_0)}{f(y_i; \widehat{\Psi})}, \qquad \forall i \in n, \tag{6}$$

where f_{nd} is composed of normal components that are designated as nondifferential. We then classify observation y_i with weight w_i to be a differential element if $\text{fdr}(y_i)/w_i \le z_0$, for some threshold value z_0.

2.5. Software. The method presented in this paper has been implemented in an R package called DIME (differential identification using mixture ensemble) and is available at http://www.stat.osu.edu/~statgen/SOFTWARE/DIME/ or http://cran.r-project.org/web/packages/DIME/index.html (CRAN).

3. Results and Discussion

3.1. Simulation Study. Our simulation was modeled after the APO AI gene expression data [21]. Let x_{ij} be the logarithm of expression level corresponding to the ith gene (observation unit) in the jth sample ($j = 1$ if it is a control sample and $j = 2$ if it is a treatment sample). For nondifferential genes, we generated $x_{i2} - x_{i1}$ by sampling randomly from genes in the APO AI dataset for which the log-ratio is at most one. For differential genes, we simulated the log of expression level in the control sample from a uniform distribution (i.e., $x_{i1} \sim \text{unif}(15, 30)$); we set the log expression level in the treatment sample to be $x_{i2} = x_{i1} + Z_{i2} + (2 \times B_{i2} - 1) \times G_{i2}$, where $Z_{i2} \sim N(0, 0.7 - 0.02 \times x_{i1})$, $B_{i2} \sim \text{Bern}(0.5)$, and G_{i2} followed one of the following three distributions:

$$G_{i2} \sim \begin{cases} (1) & \text{exponential}(\beta = 1.4286) + 1, \\ (2) & \text{uniform}(1, 4), \\ (3) & \text{normal}(\mu = 2.5, \sigma = 0.75). \end{cases} \tag{7}$$

Note that Z_{i2} was set such that genes with smaller expression will have larger variance, while B_{i2} controls over- or underexpression of genes. Further, G_{i2} represents three different underlying distributions for differential observations to study the performance of the ensemble model under diverse data types. We generated 10,000 genes for which 10% (1000) are differential elements. A total of 100 datasets were simulated under each of the three simulation settings (the three G distributions in (7)). In each replicate, we calculated false positive rate (FPR) and true positive rate (TPR) for classifying each gene as differential or nondifferential. Here, TPR is the rate of correct classification of differential genes and FPR is the rate of classifying a gene to be differential when it is in fact a nondifferential gene. Figure 1 shows the result of the ensemble approach fitting these three types of simulated data. In column 1 (datasets with exponential distribution for

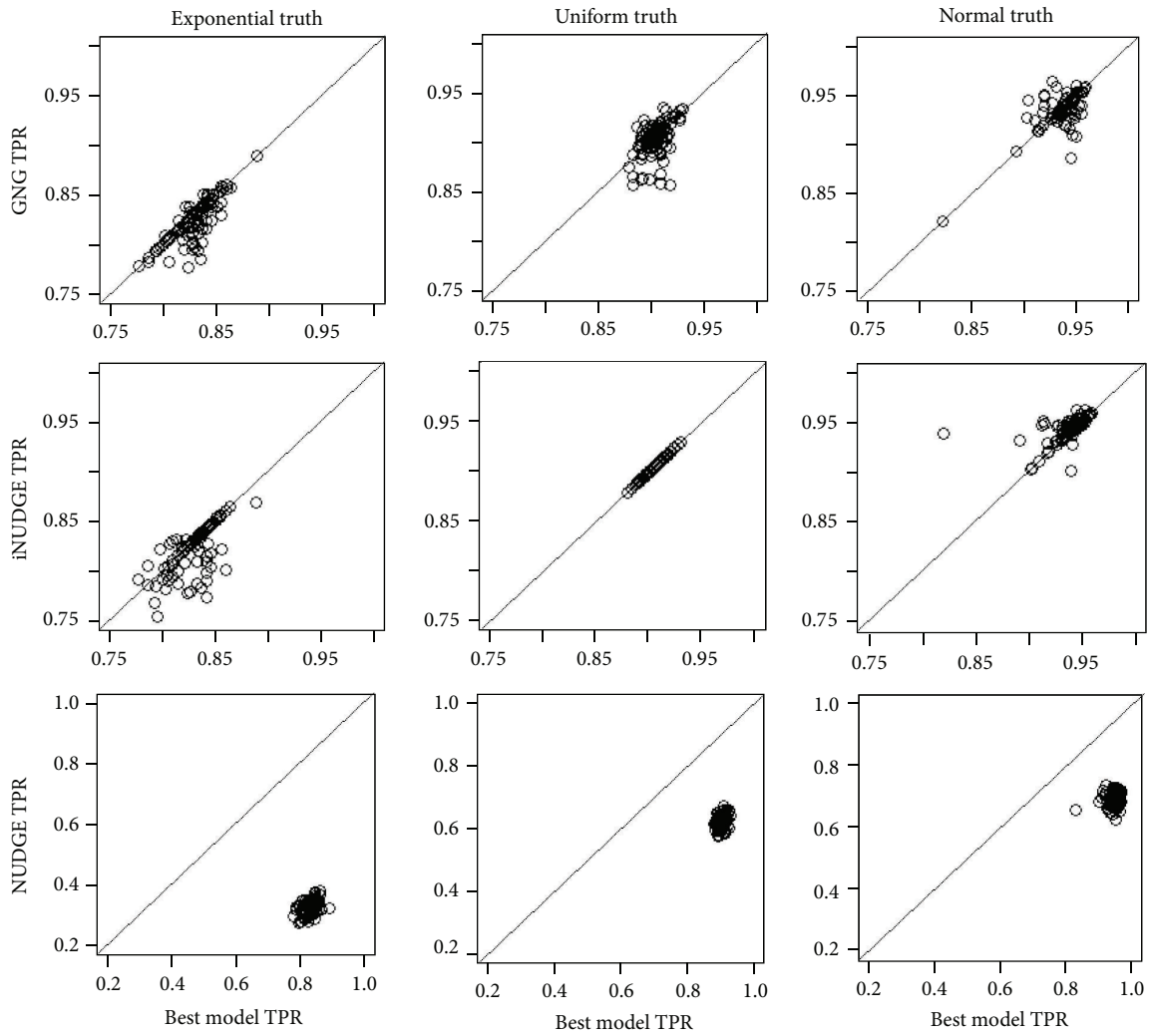


Figure 1: Simulation results comparing the performance of the ensemble approach with each of the three individual classes of models, GNG (row 1), iNUDGE (row 2), and NUDGE (row 3), under 3 different underlying models, exponential (column 1), uniform (column 2), and normal (column 3), representing three different data types.

G), we can see that the best model selected was GNG for a majority of the replicates, which is evident from the fact that the majority of the points are on the diagonal line for the best model versus GNG plot (row 1, column 1). Although iNUDGE also has similar TPR (row 2, column 1), it is more variable (more scattering) with a slightly lower average TPR. In column 2 (datasets with uniform distribution), iNUDGE was chosen as the best model in all replicates, while GNG has slightly lower TPR and more scattering (row 1, column 2), opposite of column 1. For model with underlying normal truth (3rd column), iNUDGE was selected to be the best in some cases whereas in other cases GNG was selected as the best overall model. Overall, regardless of the underlying distribution, the best model selected has comparable or better TPR compared to individual iNUDGE or GNG models. On the other hand, results from NUDGE are associated with a much lower TPR due to its limitation of using only one normal component (row 3). The false positive rates are not shown as they are zero in all replicates.

3.2. Real Data Analysis. After confirming the utility of the ensemble approach for handling multiple data types through a simulation study, we analyzed real data sets from ChIP-seq, DNA methylation, and gene expressions. ChIP-seq experiment has become the most commonly used technique to interrogate genome-wide protein-DNA interaction locations in recent years. It has enabled scientists to study transcription factor binding sites with better accuracy and less cost compared to older technology such as ChIP-chip [4]. Such data may be used to capture differential transcription factor binding sites in normal versus cancer samples, which may provide insights as to how cancer-related genes are turned on/off. For more information about ChIP-seq datasets and the methods used in preprocessing, including normalization, see Taslim et al. [11]. DNA methylation is an important factor in heritable epigenetic regulation that has been shown to alter gene expression without changes in DNA sequence. DNA methylation has been associated with many important processes such as genomic imprinting and carcinogenesis

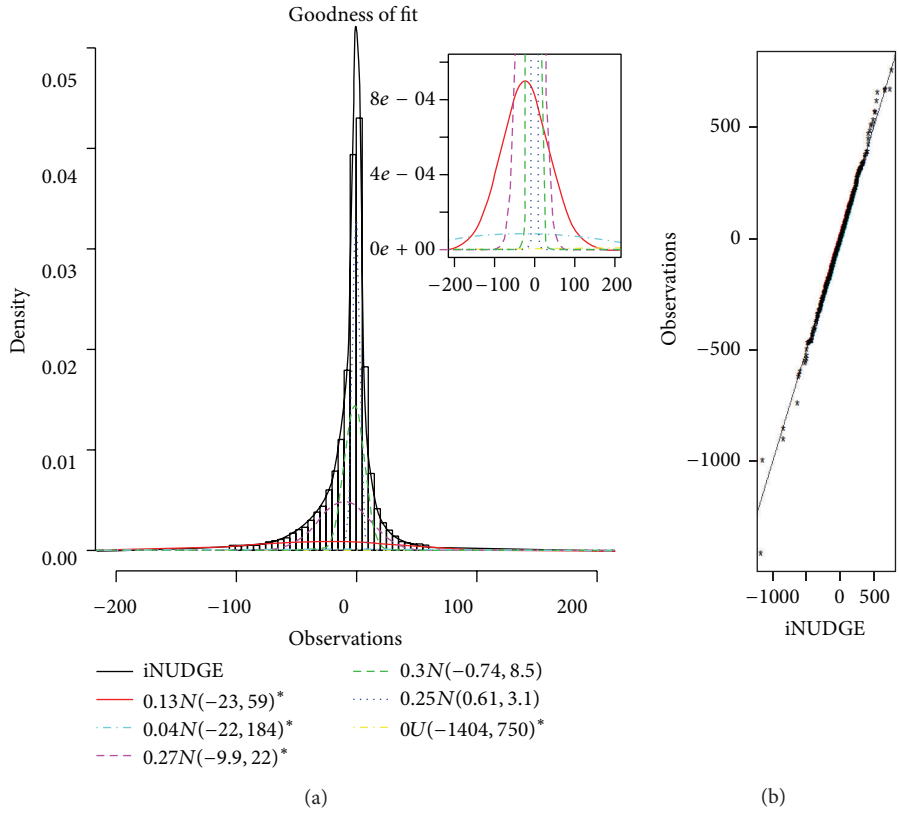


Figure 2: Results from fitting DIME to the ChIP-seq data, where ∗ designates a differential component. (a) The histogram of the methylation data is superimposed by the fitted best model and the individual components (inset: zoomed in view showing the individual components of the fitted model). (b) The QQ-plot of the best model versus the observed normalized ChIP-seq data.

Table 1: Summary of three types of real datasets.

Data type	Description	Positive control	Negative control
ChIP-seq	Pol II comparison in MCF7 versus OHT	No	No
DNA methylation	MCF7 versus pooled normal	No	No
	T47D versus pooled normal	No	No
	MDA-MB-361 versus pooled normal	No	No
Gene expression	Apo AI knocked out versus normal mice	Yes	No
	HIV infected $CD4^+$ T cell versus noninfected	Yes	Yes

[22, 23]. Likewise, differential analysis of gene expression has enabled researchers to find cancer-associated genes and other diseases [24, 25]. For more information on DNA methylation and gene expression datasets and normalization methods, see Khalili et al. [15] and Dean and Raftery [16]. Table 1 provides a summary of the three types of real data used in this section.

3.2.1. ChIP-Seq Data. The ensemble model was applied to identify genes associated with enriched polymerase II (Pol II) binding quantity in OHT (tamoxifen resistant breast cancer cell line) compared to normal breast cancer (MCF7). We used the normalized data described in Taslim et al. [11]. The ensemble modeling approach selected iNUDGE as the best overall model with 5 normal components. However, the mixing proportion for the uniform component is negligible. According to the first step of the classification criterion, three of the normal components were designated as differential components (see Figure 2(a)). Figure 2(b) shows the QQ-plot, which indicates a good fit of the model to the data. DIME identified around 21% (3,909) of the genes as having enriched Pol II binding quantity in OHT cell line when compared with MCF7.

3.2.2. DNA Methylation. The ensemble model was also applied to identify differentially methylated genes in three breast cancer cell lines: MCF7, T47D, and MDA-MB-361. These methylation data are from DMH microarrays that employed 2-color technique comparing a cancer cell line with a normal pooled DNA sample [12]. Lower Huber's weighting scheme was used to downweigh the log-ratios of

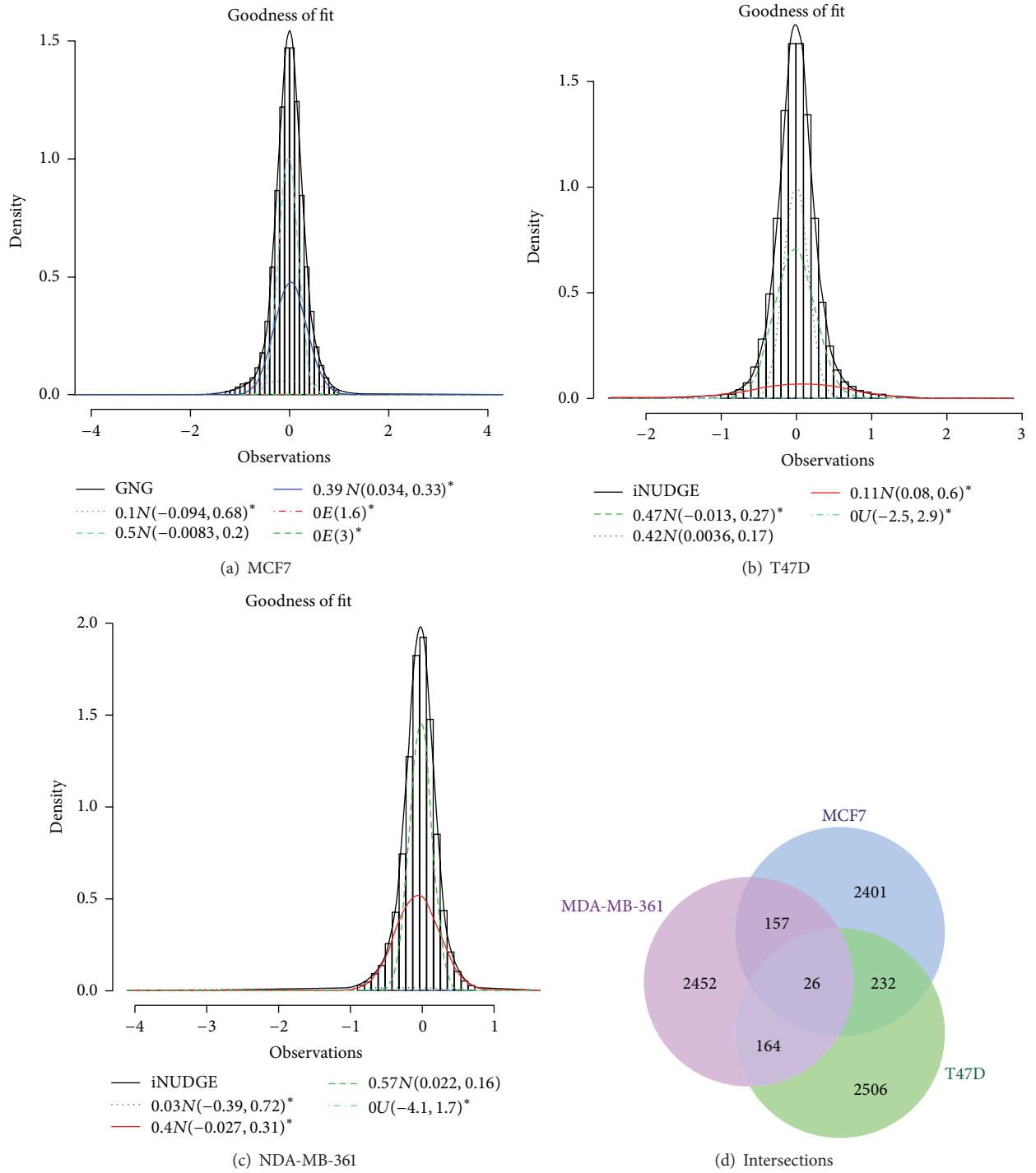


Figure 3: Results from fitting the weighted DIME to the DNA methylation data, where $*$ designates a differential component. ((a)–(c)) The histogram of the methylation data is superimposed by the fitted best model showing the fit of the model. Individual components of the best model are also shown to be superimposed on the histogram. (d) The Venn diagram showing the number of uniquely methylated loci in each of the three cell lines and the number of methylated loci shared between the three different cell lines.

small intensities. The overall best models selected for MCF7, T47D, and MDA-MB-361 cell lines were GNG, iNUDGE, and iNUDGE, respectively. Figures 3(a)–3(c) show that the model chosen can capture the distribution of all three cell lines well. It turns out that each estimated model has 3 normal components with negligible uniform or exponential. This indicates the need for normal component(s) to represent the differential probes, and indeed two of the three normal components were labeled as differential in each of the three cell lines. The number of probes identified to be differentially

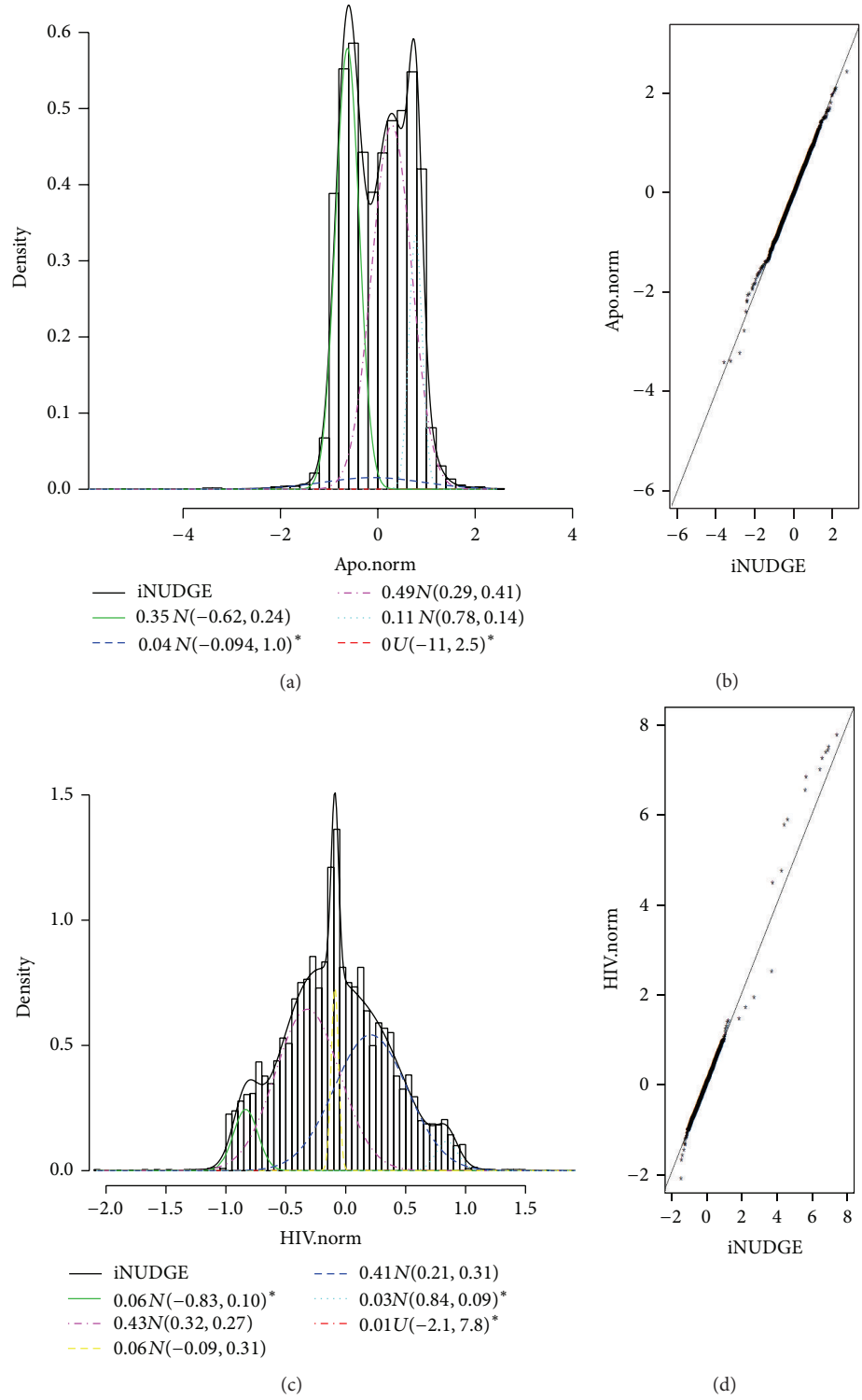


FIGURE 4: Results from fitting the DIME models to two gene expression data: ((a) and (b)) Apo AI results; ((c) and (d)) HIV results; ((a) and (c)) histograms of the normalized data superimposed by the fitted best model along with their individual components; and ((b) and (d)) QQ-plot of the fitted model versus the normalized data.

methylated is 2816, 2928, and 2799 (among around 44 k probes) for MCF7, T47D, and MDA-MB-361, respectively. The three cell lines are known to be heterogeneous and as such many uniquely methylated loci are expected. Nevertheless, since all three cell lines are hormone-receptor positive, some shared methylated loci should also be present. Indeed,

the majority of the probes are uniquely methylated in each cell line but there are a few that are shared between the different cell lines (Figure 3(d)).

3.2.3. Gene Expression. Dean and Raftery [16] analyzed a couple of data sets that have some known differentially expressed genes and some known similarly expressed genes, which made them valuable test data sets as it is possible to check for false positive and false negative rates.

Dataset I: Apo AI. In this experiment, gene expression was obtained from eight normal mice and eight mice with their APO AI gene knocked out [21]. DIME yielded iNUDGE with 4 normal components as the best overall model, which indeed capture the trimodal feature of the data well (Figure 4(a)). The goodness of fit of iNUDGE is further supported by the accompanying QQ-plot (Figure 4(b)). In this model, one of normal densities was labeled as a differential component. Based on this model, 31 genes were identified as differentially expressed, which includes the 8 positive genes discussed in Dudoit et al. [21].

Dataset III: HIV Data. This dataset compares cDNA from $CD4^+$ T cells at 1 hour after infection with HIV-1BRU and their noninfected counterparts. There were 13 genes known to be differentially expressed (HIV-1 genes, which were used as positive controls) and there were also 29 negative control genes. iNUDGE was selected once again as the best overall model for explaining the data. The density plot with 5 normal components and the QQ-plot (Figures 4(c)-4(d)) confirm the goodness of fit of the selected model. In particular, it is noted that the "spike" at the center of the distribution was quite well captured, although the QQ-plot shows disagreement between the data and the model at the right tail. There were 18 genes classified as differentially expressed, which include the 13 known positive controls. Further, none of the 29 negative control genes were included in the identified set.

4. Conclusions

Thanks to rapid progress and innovations in genomic research, scientists are now producing a great deal of diverse types of data in a very short period of time. These exciting developments however brought great challenges for carrying out appropriate data analysis. Existing methods designed specifically for a particular type may not lead to satisfactory results when applied to another data type. In this paper, we propose a unified differential identification approach based on an ensemble framework that is flexible enough to handle multitype data (from older/current technologies as well as potential future data). Our approach is based on classes of mixture model that have been proposed for specific data types. Here, we package these approaches into one unified framework synthesizing each of their individual advantages. In our proposed methodology, the best overall model will be selected depending on the underlying characteristics of the data and classification based on this model will be performed.

We demonstrated the applicability of our approach using both simulated and real data. We simulated data under three different underlying distributions to show the versatility of our methods for analysis for different types of data. Our results indeed show that the best model chosen by the program performed as well as or, in most cases, better than individual results from GNG, NUDGE, or iNUDGE, regardless of the underlying data types. Furthermore, it is clear from the simulation study that NUDGE is not a competitive model. The main reason for NUDGE's poor performance is due to the fact that it does not have the ability to adapt to different data types without the multiple normal components. Having multiple normal components in GNG and iNUDGE is shown to be essential to capture nondifferential elements that is not symmetrical and sometimes may even be multimodal. In our approach, labeling some normal components as differential turned out to be beneficial in that it allows the best model to capture differential data coming from any distribution, thereby increasing the flexibility of our ensemble model to capture diverse data types. Results from the analysis of three real data types all lead to reasonable goodness of fit. Further, good classification power and low error rates were obtained when applied to data with known positive and known negative controls.

Our model uses mixture of normal components with unequal variances, which can lead to a singularity problem (unboundedness in likelihood when a component variance is 0) in some cases. One suggestion to alleviate this problem to some extent is to use the BIC model selection criterion to discourage larger model (hence, less chance of having a component with observations all having the same value), which we have implemented in the package. One may also use a clustering algorithm (e.g., K-means) to provide reasonable initial starting parameters for the mixture model. In our implementation, we display a warning if potential singularity is detected. Restarting the model from different initial values and/or random seeds would also be recommended. In fact, as a good practice, our model should be run in many iterations with different starting parameters to avoid simply finding local optimum. A penalized likelihood may also be entertained to steer the variance estimates away from zero [26]. In our approach, we designate normal component as capturing differential regions based on what is commonly perceived as extreme values. Thus, the result of the classification may vary if this cut-off value is set differently.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported in part by the National Cancer Institute (Grant U54CA113001) and the National Science Foundation (Grant DMS-1042946).

References

[1] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a

complementary DNA microarray," *Science*, vol. 270, no. 5235, pp. 467–470, 1995.

[2] Q. Wang, W. Li, Y. Zhang et al., "Androgen receptor regulates a distinct transcription program in androgen-independent prostate cancer," *Cell*, vol. 138, no. 2, pp. 245–256, 2009.

[3] P. Bertone, V. Stolc, T. E. Royce et al., "Global identification of human transcribed sequences with genome tiling arrays," *Science*, vol. 306, no. 5705, pp. 2242–2246, 2004.

[4] D. S. Johnson, A. Mortazavi, R. Myers, and B. Wold, "Genome-wide mapping of in vivo protein-DNA interactions," *Science*, vol. 316, no. 5830, pp. 1497–1502, 2007.

[5] K. D. Hansen, B. Langmead, and R. A. Irizarry, "BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions," *Genome Biology*, vol. 13, no. 10, article R83, 2012.

[6] T. H.-M. Huang, M. R. Perry, and D. E. Laux, "Methylation profiling of CpG islands in human breast cancer cells," *Human Molecular Genetics*, vol. 8, no. 3, pp. 459–470, 1999.

[7] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biology*, vol. 11, no. 10, article R106, 2010.

[8] D. J. McCarthy, Y. Chen, and G. K. Smyth, "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation," *Nucleic Acids Research*, vol. 40, no. 10, pp. 4288–4297, 2012.

[9] Z. Ouyang, Q. Zhou, and W. H. Wong, "ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 51, pp. 21521–21526, 2009.

[10] G. Robertson, M. Hirst, M. Bainbridge et al., "Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing," *Nature Methods*, vol. 4, no. 8, pp. 651–657, 2007.

[11] C. Taslim, J. Wu, P. Yan et al., "Comparative study on ChIP-seq data: normalization and binding pattern characterization," *Bioinformatics*, vol. 25, no. 18, pp. 2334–2340, 2009.

[12] M. Weber, J. J. Davies, D. Wittig et al., "Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells," *Nature Genetics*, vol. 37, no. 8, pp. 853–862, 2005.

[13] E. Pettersson, J. Lundeberg, and A. Ahmadian, "Generations of sequencing technologies," *Genomics*, vol. 93, no. 2, pp. 105–111, 2009.

[14] N. Rusk, "Torrents of sequence," *Nature Methods*, vol. 8, no. 1, article 44, 2011.

[15] A. Khalili, T. Huang, and S. Lin, "A robust unified approach to analyzing methylation and gene expression data," *Computational Statistics & Data Analysis*, vol. 53, no. 5, pp. 1701–1710, 2009.

[16] N. Dean and A. E. Raftery, "Normal uniform mixture differential gene expression detection for cDNA microarrays," *BMC Bioinformatics*, vol. 6, article 173, 2005.

[17] P. J. Huber, *Robust Statistics*, John Wiley & Sons, New York, NY, USA, 1981.

[18] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[19] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2–8, 1971*, pp. 267–281, 1973.

[20] B. Efron, "Large-scale simultaneous hypothesis testing: the choice of a null hypothesis," *Journal of the American Statistical Association*, vol. 99, no. 465, pp. 96–104, 2004.

[21] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed, "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments," *Statistica Sinica*, vol. 12, no. 1, pp. 111–139, 2002.

[22] J. Craig and N. Wong, Eds., *Epigenetics: A Reference Manual*, Caister Academic Press, 2011.

[23] T. Zuo, T.-M. Liu, X. Lan et al., "Epigenetic silencing mediated through activated PI3K/AKT signaling in breast cancer," *Cancer Research*, vol. 71, no. 5, pp. 1752–1762, 2011.

[24] V. Bourdeau, J. Deschênes, D. Laperrière, M. Aid, J. H. White, and S. Mader, "Mechanisms of primary and secondary estrogen target gene regulation in breast cancer cells," *Nucleic Acids Research*, vol. 36, no. 1, pp. 76–93, 2008.

[25] M. Fan, P. S. Yan, C. Hartman-Frey et al., "Diverse gene expression and DNA methylation profiles correlate with differential adaptation of breast cancer cells to the antiestrogens tamoxifen and fulvestrant," *Cancer Research*, vol. 66, no. 24, pp. 11954–11966, 2006.

[26] J. Chen, X. Tan, and R. Zhang, "Inference for normal mixtures in mean and variance," *Statistica Sinica*, vol. 18, no. 2, pp. 443–465, 2008.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 813789, 12 pages
http://dx.doi.org/10.1155/2014/813789

Research Article

Leaky Vaccines Protect Highly Exposed Recipients at a Lower Rate: Implications for Vaccine Efficacy Estimation and Sieve Analysis

Paul T. Edlefsen

Fred Hutchinson Cancer Research Center, Seattle, WA 91809, USA

Correspondence should be addressed to Paul T. Edlefsen; pedlefse@fhcrc.org

Received 31 January 2014; Accepted 14 March 2014; Published 7 May 2014

Academic Editor: Samsiddhi Bhattacharjee

"Leaky" vaccines are those for which vaccine-induced protection reduces infection rates on a per-exposure basis, as opposed to "all-or-none" vaccines, which reduce infection rates to zero for some fraction of subjects, independent of the number of exposures. Leaky vaccines therefore protect subjects with fewer exposures at a higher effective rate than subjects with more exposures. This simple observation has serious implications for analysis methodologies that rely on the assumption that the vaccine effect is homogeneous across subjects. We argue and show through examples that this heterogeneous vaccine effect leads to a violation of the proportional hazards assumption, to incomparability of infected cases across treatment groups, and to nonindependence of the distributions of the competing failure processes in a competing risks setting. We discuss implications for vaccine efficacy estimation, correlates of protection analysis, and mark-specific efficacy analysis (also known as sieve analysis).

1. Introduction

Public health vaccines have reduced the global burden of disease considerably over the past century. Statistical design and analysis of vaccine efficacy trials are well-studied and critical components of the development of these interventions. As discussed in [1], analysis of vaccine interventions is usually complicated by the unobservability of exposure. Even when exposure rates are constant across subjects, the stochastic nature of exposures means that some subjects will experience no exposures while others may experience multiple exposures. Except in challenge trials in which exposure is controlled by the experimental setting, or in controlled scenarios in which exposure is estimable; the missingness of exposure times poses a challenge to estimation of per-exposure vaccine efficacy.

Vaccine efficacy has multiple definitions (see [2] for a thorough review), including per-exposure reduction in susceptibility, which is distinct from reduction in instantaneous hazard of infection and also from reduction in overall (attack) rate of infection. These definitions coincide in some settings but generally are not the same. It has been shown that the mechanism of the vaccine's protection is relevant to the relationship among these kinds of efficacy, with "leaky" vaccines (defined as those modifying per-exposure infection rates for all subjects equally) at one extreme and "all-or-none" vaccines (which completely protect some subjects and have no effect on the others) at the other extreme. While for all-or-none vaccines the overall attack rate is reduced by the fraction of recipients that have protective responses, for leaky vaccines the attack rate is reduced by an amount that depends on the number of exposures that each subject experiences.

If each subject experiences exactly one exposure during the trial, then a leaky vaccine reducing susceptibility by 50% has the same attack-rate efficacy as an all-or-none vaccine that fully protects 50% of the subjects. Here we focus on examples such as HIV-1 vaccine trials, in which multiple exposures are possible and in which some (or many) participants will experience no exposures at all. In such settings, the effect of a partially efficacious leaky vaccine is to reduce attack rates for subjects who experience one exposure more than for subjects who experience multiple exposures, since each exposure has an independent opportunity to infect. Although in this setting reinfection is possible, we assume that the

endpoint of interest is initial infection only, so that infected subjects are removed from the at-risk population.

In this paper, we consider the analysis of leaky vaccines when there is heterogeneity in subjects' infecting exposure distributions (defined as either heterogeneity in exposure or in per-exposure infection susceptibility or both). Through arguments and simulation, [3] have previously shown that, for this scenario, the assumption of proportional hazards (that is usually required for Cox modeling) is violated. Here we restate these arguments and consider additional implications for survival analysis in the setting of competing risks. We argue that the conditional distribution of exposure rates, given infection status, depends on both time and treatment assignment, which implies not only that the hazard ratio varies over time (reflecting variation in the risk group distribution among the "at-risk" uninfected population) but also that the risk group distribution varies among those infected—both over time and across treatment groups. We discuss general implications of this observation for vaccine efficacy analysis methods and for immune correlates analysis methods, including case-only methods (which save resources by evaluating covariates only among the subjects who became infected in a trial), and for competing risks settings. We review proofs for two mark-specific efficacy analysis (also called "sieve analysis") methodologies and show that the proofs do not apply in this setting, leading to a potential bias in these analyses. We conclude that in the absence of exposure data, failure time and failure type data alone are insufficient to distinguish per-exposure vaccine efficacy that varies across subjects from per-exposure vaccine efficacy that varies across marks of the failure.

2. Materials and Methods

2.1. Notation and Definitions. In this section, we introduce the notation and examples that we will use to demonstrate the implications of risk heterogeneity for evaluating the efficacy of leaky vaccines. We assume a setting of a well-conducted placebo-controlled randomized clinical trial to evaluate a vaccine intervention, where the effect of the intervention is to reduce the per-exposure infection probability by a (multiplicative) factor η, so that if for a subject the probability of infection given one exposure is ϕ_p in the absence of the intervention, it is $\phi_v = \eta\phi_p$ if the subject is assigned to the vaccine treatment group.

As shown in [1], for nonharmful vaccines, the vaccine effect can be seen as a filter on each subject's infecting exposure process N, which is itself a filtered version of the exposure counting process E. That is, for an arbitrary process $E(t)$ counting a placebo recipient's exposures up to time t, an infection occurs with probability ϕ_p for each time t at which the exposure count increases. For vaccine recipients this probability is reduced to $\phi_v = \phi_p\eta$, where with probability $1 - \eta$ the would-be-infection is avoided due to the vaccine intervention. With minor adjustments the arguments in this paper can be adapted to apply to vaccines that could induce harm, such that the vaccine is not providing an additional filter but is modifying and possibly increasing the rate at which exposures become infections; for simplicity of presentation we will proceed with the assumption that $0 < \eta < 1$.

We assume that we do not observe the exposure processes at all; we are given data of the form of per-subject pairs (T, M) representing the observed part of the latent pair of processes (I, C), where I is the time at which the subject's infection count N increases from zero to one and C is the right-censoring time. We only observe one value of this latent pair, $T \equiv \min(I, C)$. $M = 1$ indicates missingness of the infection time ($M = 1$ means that $T = C$). We assume conditions of noninformative censoring, such that $I \perp\!\!\!\perp C$. The arguments are easily extended to a setting in which the right-censored values are used to improve estimates of efficacy, but henceforth we consider only the uncensored data (except when explicitly addressing the assumption in the context of competing risks analysis).

We define three distinct notions of vaccine efficacy, based on different quantities. First we define the attack rate for treatment group x (vaccine recipients have $x = v$ and placebo recipients have $x = p$) as $a_x = \Pr(T < \tau \mid x)$. Then the attack-rate vaccine efficacy $\mathrm{VE}^a = 1 - a_v/a_p$ is the reduction in the total fraction of infected subjects due to the vaccine. The per-exposure vaccine efficacy $\mathrm{VE}^\phi = 1 - \phi_v/\phi_p$ is the reduction in the per-exposure susceptibility to infection due to the vaccine. Finally we define the hazard-rate vaccine efficacy $\mathrm{VE}^\lambda = 1 - \lambda_v(\tau)/\lambda_p(\tau)$, where for each treatment group x the infection hazard is $\lambda_x(t) = \lim_{d \searrow 0} \Pr(N(t + d) = 1 \mid x, N(t) = 0)/d$, the instantaneous rate of infection just after time t given noninfection up to time t. The set of subjects with treatment assignment x that are not infected up to time t is called the at-risk group $R_x(t)$, and the set of subjects already infected by time t is called the infected group $I_x(t)$.

2.2. Risk Groups. We assume for simplicity of presentation that there are two risk groups. We allow that some fraction π_h of subjects is "high risk," by which we mean that the exposure rates are higher for these subjects, and in particular that both single and multiple exposures are more likely for these subjects. Since a "leaky" vaccine only protects a subject if every exposure is noninfecting, the attack-rate VE is higher for low-risk subjects than for high-risk subjects. For example, if the vaccine effect reduces per-exposure susceptibility by $\eta = 50\%$, and if low-risk subjects tend to have about one exposure during the trial and high-risk subjects tend to have about nine exposures, then about half of the low-risk subjects will be protected while about $0.5^9 = 0.2\%$ of the high-risk subjects will be protected. This implies that the fraction of high- (versus low-) risk subjects among the infected vaccine recipients will differ from that fraction among the infected placebo recipients.

For illustration, we suppose arbitrarily that the baseline hazard function is constant, as in the exponential model. Under this model, an infection event occurs in a low-risk placebo recipient at a time-constant rate λ_l, which can be written as the low-risk marginal rate of an exposure λ_l^E times the conditional probability ϕ_{lp} that the exposure will

infect the low-risk placebo recipient: $\lambda_l \equiv \lambda_l^E \phi_{lp}$. For high-risk subjects we have corresponding infecting exposure rate $\lambda_h \equiv \lambda_h^E \phi_{hp}$. Because of the memoryless property of the exponential model, these rates are equivalently viewed as hazards.

The fraction a_{lp} of low-risk placebo recipients that will become infected is the fraction having infecting exposure times T that exceed the trial duration τ. Since we assume independence across subjects, under the exponential model the number of infected low-risk placebo recipients follows a binomial distribution with proportion a_{lp} given by the probability that a Poisson-distributed random variable (with rate $\lambda_l \tau$) counting infecting exposures exceeds zero: $a_{lp} \equiv 1 - e^{-\lambda_l \tau}$. Similarly the number of the high-risk placebo recipients that will become infected is given by a binomial distribution with proportion $a_{hp} \equiv 1 - e^{-\lambda_h \tau}$.

2.3. VE^a under Heterogeneous Risk. A leaky vaccine with multiplicative vaccine effect $\eta = \phi_{lv}/\phi_{lp} = \phi_{hv}/\phi_{hp}$ (corresponding to a VE$^\phi$ of $1 - \eta$) will result in overall attack rates $a_{lv} = 1 - e^{-\lambda_l \tau \eta}$ and $a_{hv} = 1 - e^{-\lambda_h \tau \eta}$ among low- and high-risk vaccine recipients, respectively. If the vaccine is partially efficacious then $0 < \eta < 1$, and $a_{hv} < a_{hp}$ and $a_{lv} < a_{lp}$, so the vaccine reduces the probability of being infected for both high- and low-risk participants. However, the reduction is not the same for high-risk participants as for low-risk participants, since $\lambda_h > \lambda_l$ implies that

$$\left(\frac{a_{hv}}{a_{hp}} = \frac{1 - e^{-\lambda_h \eta \tau}}{1 - e^{-\lambda_h \tau}}\right) > \left(\frac{a_{lv}}{a_{lp}} = \frac{1 - e^{-\lambda_l \eta \tau}}{1 - e^{-\lambda_l \tau}}\right). \tag{1}$$

The direction of the inequality is reversed for harmful vaccines (with $\eta > 1$).

2.4. Differential Enrichment of High-Risk Infected Subjects across Treatment Groups. This differential attack-rate efficacy by risk group results in a different proportion of high-risk participants among infected subjects at the end of the trial across the two treatment groups. To see this, consider that if the beginning-of-trial probability of being high risk is π_h, then we can define the conditional probability γ_{hx} of being high risk for subjects in the infected group $I_x(\tau)$ in terms of the posterior odds $\gamma_{hp}/(1 - \gamma_{hp}) \equiv (\pi_h/(1 - \pi_h))(a_{hp}/a_{lp})$ for placebo recipients and $\gamma_{hv}/(1 - \gamma_{hv}) \equiv (\pi_h/(1 - \pi_h))(a_{hv}/a_{lv})$ for vaccinees.

For partially efficacious vaccines with $0 < \tau < 1$, since the vaccine reduces low-risk infections more than high-risk infections, $a_{lv}/a_{lp} < a_{hv}/a_{hp}$. This results in an enrichment of high-risk participants among the infected vaccinees as compared with the infected placebo recipients: $\gamma_{hv} > \gamma_{hp}$. For a harmful vaccine, this inequality is reversed.

2.5. Differential Enrichment of High-Risk at-Risk Subjects across Treatment Groups. This correspondingly results in a different proportion of high-risk participants among subjects remaining at-risk at the end of the trial across the two treatment groups. The posterior odds of being high risk among those remaining uninfected are $\omega_{hp}/(1 - \omega_{hp}) \equiv (\pi_h/(1 - \pi_h))((1 - a_{hp})/(1 - a_{lp}))$ for placebo recipients and $\omega_{hv}/(1 - \omega_{hv}) \equiv (\pi_h/(1 - \pi_h))((1 - a_{hv})/(1 - a_{lv}))$ for vaccinees.

If $(1 - a_{hp})/(1 - a_{lp}) < (1 - a_{hv})/(1 - a_{lv})$, or equivalently if $(1 - a_{lv})/(1 - a_{lp}) < (1 - a_{hv})/(1 - a_{hp})$, then this results in an enrichment of high-risk participants among the uninfected vaccinees as compared with the uninfected placebo recipients: $\omega_{hv} > \omega_{hp}$. This condition is met if both $a_{lv}/a_{lp} < a_{hv}/a_{hp}$ and $(a_{hp} - a_{hv}) > (a_{lp} - a_{lv})$, since we can write

$$\begin{aligned} &\frac{1 - a_{lv}}{1 - a_{lp}} < \frac{1 - a_{hv}}{1 - a_{hp}} \\ &\text{as } a_{lv}a_{hp} + \left(a_{lp} - a_{lv}\right) < a_{hv}a_{lp} + \left(a_{hp} - a_{hv}\right). \end{aligned} \tag{2}$$

For a partially efficacious vaccine we have shown that $a_{lv}/a_{lp} < a_{hv}/a_{hp}$, which implies that $a_{lv}a_{hp} < a_{hv}a_{lp}$, so if also $(a_{hp} - a_{hv}) > (a_{lp} - a_{lv})$, then the condition in (2) is satisfied.

We may still have $\omega_{hv} > \omega_{hp}$ despite not satisfying $(a_{hp} - a_{hv}) > (a_{lp} - a_{lv})$ and $a_{lv}/a_{lp} < a_{hv}/a_{hp}$. The general condition is that

$$a_{hv}a_{lp} - a_{lv}a_{hp} > \left(a_{lp} - a_{lv}\right) - \left(a_{hp} - a_{hv}\right). \tag{3}$$

2.6. Summary. In this section we have shown that, for leaky vaccines, subject heterogeneity in risk results in time variation of VE$^a = 1 - a_v/a_p$, where the values a_x (for $x \in \{v, p\}$) are the marginal attack rates for vaccine and placebo recipients. We have shown that this implies a change in the composition of both the infected group $I_x(t)$ and in the at-risk group $R_x(t)$ over time such that for both vaccine and placebo recipients the proportion of high-risk subjects is higher in the infected group than in the at-risk group by the end of the trial. We have shown that this effect differs by treatment group such that for partially protective leaky vaccines, the fraction γ_{hv} of high-risk subjects among those infected in the vaccine group is higher than the fraction γ_{hp} of high-risk subjects among those infected in the placebo group. Finally we have shown that the proportion of high-risk subjects among those remaining at-risk at the end of the trial may be higher or lower in the vaccine group as compared with the placebo group; the crucial point is that in general one should not expect that the at-risk groups have the same distribution of high-risk subjects across treatments arms.

3. Results and Discussion

Next, we turn to implications of these observations. First, we show, as has been shown previously, that VE$^\lambda$ changes over time or equivalently that the hazard proportion is inconstant. Then we discuss implications of the risk imbalance in the infected group for introducing bias into correlates of protection analysis whenever a putative correlate of protection is also a correlate of placebo-recipient risk. Finally, we discuss implications of the risk imbalance in the at-risk group in a competing risks analysis and show that this risk imbalance violates conditions required for the correctness of proofs of unbiasedness for two sieve analysis methods for leaky

vaccines, with the implication that the proven unbiasedness is only guaranteed if subject risk is homogeneous.

3.1. Implications for the Proportional Hazards Assumption. The differential efficacy for high-risk and low-risk subjects has the effect of inducing a violation of the proportional hazards assumption for the marginal hazards, even if it holds separately for the low-risk hazards and the high-risk hazards. Each marginal hazard function is a mixture of the two risk-group hazards, and the mixing proportion changes over time differently for placebo recipients than for vaccine recipients as the at-risk frequencies diverge due to different rates of infection in the two risk groups.

The marginal hazard rate of infection is a mixture over high- and low-risk subjects. At the beginning of the trial the marginal hazard for placebo recipients is $\lambda_p(0) \equiv \pi_h\lambda_h + (1-\pi_h)\lambda_l$. This changes over the trial, since $\lambda_p(\tau) \equiv \omega_{hp}\lambda_h + (1-\omega_{hp})\lambda_l$. The change is due to a shifting mixing proportion, and it appears even when there are constant hazards within each risk group.

For vaccine recipients, there is also a change in the marginal hazard over the course of the trial, but the change is different than for placebo recipients. At the beginning of the trial the marginal hazard for vaccine recipients is $\lambda_v(0) \equiv \pi_h\lambda_h\eta + (1-\pi_h)\lambda_l\eta$. At the end of the trial, $\lambda_v(\tau) \equiv \omega_{hv}\lambda_h\eta + (1-\omega_{hv})\lambda_l\eta$.

If the study enrolls n vaccine recipients, $n * \pi_h$ of whom are high risk, then a a_{l_v} infection rate among low-risk vaccine recipients (and a corresponding a_{h_v} among the high-risk vaccinees) over the course of the trial yields a difference in the ratio of high : low risk at-risk subjects from π_h at the beginning to ω_{hv} at the end. Since the high-risk hazard rate is λ_h/λ_l times the low-risk hazard rate λ_l, then the marginal hazard goes from $((1-\pi_h)\lambda_l\eta + \pi_h(\lambda_h/\lambda_l)\lambda_l\eta) = (1-\pi_h + \pi_h(\lambda_h/\lambda_l))\lambda_l\eta$ to $((1-\omega_{hv})\lambda_l\eta + \omega_{hv}(\lambda_h/\lambda_l)\lambda_l\eta) = (1-\omega_{hv} + \omega_{hv}(\lambda_h/\lambda_l))\lambda_l\eta$. The vaccine recipient hazard is $1-(1-\omega_{hv} + \omega_{hv}(\lambda_h/\lambda_l))/(1-\pi_h + \pi_h(\lambda_h/\lambda_l))$ times 100% lower at the end of the trial than at the beginning. The placebo recipient hazard is correspondingly $1-(1-\omega_{hp}+\omega_{hp}(\lambda_h/\lambda_l))/(1-\pi_h + \pi_h(\lambda_h/\lambda_l))$ times 100% lower at the end of the trial than at the beginning.

Unless the end-of-trial rates of high-risk subjects among the uninfected are the same for both treatment groups (i.e., unless $\omega_{hp} = \omega_{hv}$), the hazard ratio (vaccine to placebo) will also differ at the end of the trial. The marginal hazards ratio at the beginning of the trial is $\lambda_v(0)/\lambda_p(0) = (\pi_h\lambda_h\eta + (1-\pi_h)\lambda_l\eta)/(\pi_h\lambda_h + (1-\pi_h)\lambda_l)$. At the end of the trial it is $\lambda_v(\tau)/\lambda_p(\tau) = (\omega_{hv}\lambda_h\eta + (1-\omega_{hv})\lambda_l\eta)/(\omega_{hp}\lambda_h + (1-\omega_{hp})\lambda_l)$. These are equal only when $\omega_{hv} = \omega_{hp} = \pi_h$, and never for leaky vaccines with heterogeneous risk.

We demonstrate the situation with a simple example of a leaky vaccine with about $a_{l_p} = 4\%$ of low-risk placebo recipients becoming infected over the unit-time course of the trial (corresponding to a low-risk infecting exposure rate of $\lambda_l = 0.04$) and about $a_{h_p} = 36\%$ of high-risk placebo recipients becoming infected (corresponding to a high-risk instantaneous infecting exposure rate of $\lambda_h = 0.446$). We suppose a leaky vaccine that reduces the infection probability by $\eta = 50\%$ per exposure, which corresponds to $a_{l_v} = 2\%$ of low-risk vaccinees and $a_{h_v} = 20\%$ of high-risk vaccinees becoming infected over the course of the trial. We suppose that, at the start of the trial, $\pi_h = 5\%$ of participants are high risk.

If the study enrolls 100 vaccine recipients, 5 of whom are high risk, then a 2% infection rate among low-risk vaccine recipients (and a corresponding 20% among the high-risk vaccinees) over the course of the trial yields a difference in the mixture of high : low risk hazards from 5 : 95 ($\pi_h = 5\%$) at the beginning to 4 : 93 ($\omega_{hv} = 4.1\%$). Since in our example the high-risk hazard rate λ_h is about eleven times the low-risk hazard rate λ_l, then the marginal vaccine hazard goes from $(0.95 \times \lambda_l\eta + 0.05 \times 11\lambda_l\eta) = 1.5\lambda_l\eta$ to $(0.96 \times \lambda_l\eta + 0.04 \times 11\lambda_l\eta) = 1.4\lambda_l\eta$. In this example, the vaccine recipient marginal hazard is about 5.8% lower at the end of the trial than at the beginning.

If that study also enrolls 100 placebo recipients, 5 of whom are high risk, then a 4% infection rate among low-risk placebo recipients and a corresponding 36% among the high-risk placebos over the course of the trial yields a difference in the mixture of high : low risk hazards from 5 : 95 at the beginning to 3 : 91 (about $\omega_{hp} = 3.4\%$). Then the marginal placebo hazard goes from $(0.95 \times \lambda_l + 0.05 \times 11\lambda_l) = 1.5\lambda_l$ to $(0.976 \times \lambda_l + 0.034 \times 11\lambda_l) = 1.34\lambda_l$. In this example, the marginal placebo recipient hazard is about 10.7% lower at the end of the trial than at the beginning.

For the conditions of our example, the hazard ratio at the beginning of the trial (vaccine/placebo) is $1.5\eta/1.5 = 0.5$, but at the end of the trial it is $1.4\eta/1.34 = 0.527$, about a 5.5% increase. If we increase the rate of exposures for high-risk subjects λ_h to 1, so that we expect about one exposure per high-risk participant, then the ending hazard ratio is about 18.9% higher than it is at the trial's beginning. The discrepancy peaks at about $\lambda_h = 3$, at a 55% increase in the hazard ratio, then decreases again as the leaky vaccine effect diminishes for the high-risk subjects. At $\lambda_h = 10$, the ending hazard proportion is down to 8.8% above its starting value.

The hazard proportion also changes over the duration of the trial; Figure 1 shows the change over time of the hazard ratio. The plot shows that the change is nonmonotonic in time and that it has a single mode and a right skew. For harmful vaccines with $\eta > 1$, the plot has the same shape, but the plot is mirrored over the X axis, with negative percent change values indicating that the hazard ratio decreases and then increases again.

3.2. Implications for Correlates Analysis. The differential enrichment of high-risk subjects among those infected across treatment groups implies that even if a vaccine has an equal per-exposure effect on every subject, its effects on overall attack rates are expected to differ by risk group. When evaluating a vaccine candidate to determine if its partial efficacy can be attributed to unequal vaccine effects across subjects (by for instance identifying preexisting subject traits or immune responses to vaccination that differentiate subjects for whom the vaccine worked best), care must be taken to differentiate between these expected attack-rate effects (which do not reflect differential per-exposure efficacy

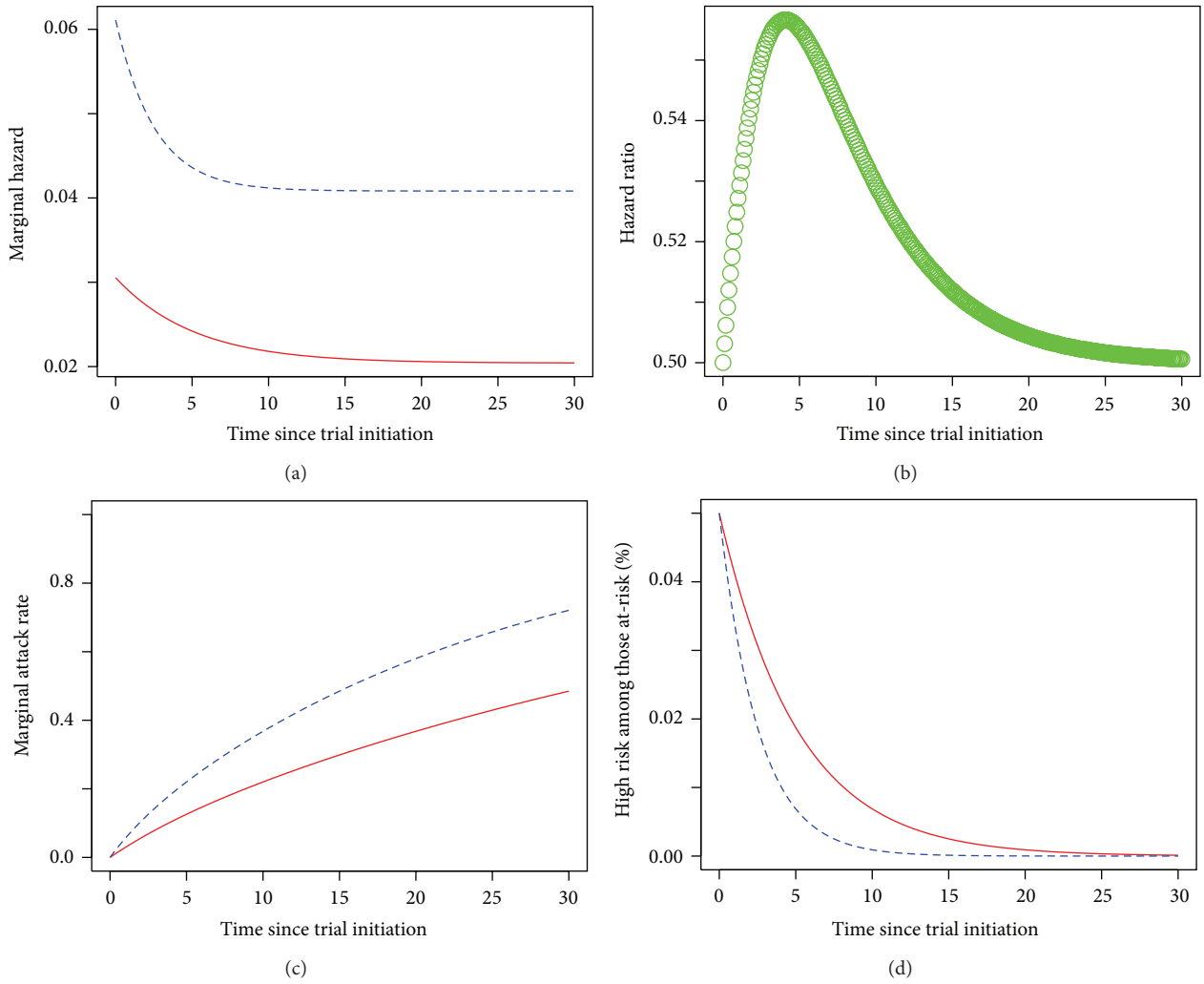


FIGURE 1: Effects of differential enrichment for high-risk subjects in the at-risk population across treatment groups. (a) The marginal hazards as a function of time for placebo recipients (dashed blue line) and vaccine recipients (solid red line) for the conditions of our example trial in which a 1 : 1 randomization allocates subjects to receive a placebo or a leaky vaccine with per-exposure efficacy $\eta = 0.5$, with independent Poisson exposure rates $\lambda_l = 0.0408$ and $\lambda_h = 0.4463$ for low-risk and high-risk subjects, respectively, and a $\pi_h = 5\%$ starting fraction of high-risk subjects. (b) The ratio of the marginal hazards in (a). (c) The fraction of subjects infected in the two groups over time. (d) The proportions $\omega_{hp}(t)$ and $\omega_{hv}(t)$ of the at-risk groups $R_p(t)$ (dashed blue line) and $R_v(t)$ (solid red line) that are high risk, over time.

by subject trait) from effects that truly modify the per-exposure efficacy by subject trait.

Several authors have noted that the analysis of vaccine trials to identify subject correlates of VE^{ϕ} is complicated by missingness of the counterfactual effects of vaccination on the placebo recipients (see [4] for a review and unifying perspective). With the data typically available from a clinical trial it is possible to estimate correlates of infection risk within vaccine recipients and placebo recipients separately but without strong assumptions or additional data it is not possible to causally attribute changes in infection risk (for some subset of subject covariates) to the vaccine treatment assignment. The problem is that it is not possible to differentiate between preexisting risk differences and vaccine-induced risk differences without additional information.

Here we point out that leaky vaccines with heterogeneous subject risk constitute a concrete example of this difficulty. Since we expect differential enrichment of high-risk subjects even when the vaccine has an equal per-exposure effect, then any correlate of infection risk in the placebo group will necessarily correlate with VE^{a}. We also expect a correlation between a subject's risk category and VE^{λ}. The implication is that (in the absence of additional justification) any identified correlate of risk in the vaccine group should not be interpreted as a correlate of protection if it is also a correlate of risk in the placebo group.

This suggests a test for any putative candidate correlate of VE^{ϕ}: if an association exists between the correlate and infection risk in the placebo group then any correlation observed in the vaccine group (even a much stronger correlation) may

be solely attributable to expected risk group enrichment and should not (without further justification) be attributed to differential per-exposure efficacy. If there is no association in the placebo group then the correlate of vaccine-group infection risk remains a plausible candidate as a correlate of VE^{ϕ}. Further work is required to develop the conditions under which a vaccine-group correlate of infection risk can be attributed to differential efficacy, but this argument suggests general caution whenever a placebo-group correlation cannot be ruled out.

This reasoning also warrants caution about so-called "case-only" methods, which evaluate only the infected cases. Such methods can be cost saving because correlates need not be measured in uninfected subjects. However if there is enrichment of different risk groups among infected subjects in the two treatment groups (which should be expected for any leaky vaccine), then covariate differences across treatment groups among infected subjects may simply reflect differential baseline risk. Since correlate information is unavailable for uninfected subjects, in case-only analyses the test of placebo-recipient risk correlation is not possible. Below we examine a special case of case-only analysis in the setting of competing risks, known as "sieve analysis."

3.3. Implications for Competing Risks and Sieve Analysis. In addition to evaluating vaccine efficacy as a function of subject-specific covariates, it is often of interest to evaluate the extent to which a vaccine's efficacy differs by type of infection. In a series of papers on what has variously been called "mark-specific intervention efficacy" or "sieve effects," Gilbert et al. defined sufficient conditions under which estimates are unbiased for quantities relevant to the identification of these effects [5–7]. Here we argue that one of those conditions can be represented as a requirement of "proportional exposure pseudohazards" and that this condition is required not only for the failure-type-only methods (such as multinomial logistic regression (MLR)) but also for the time-to-event methods (including competing risks Cox models, even when relaxing the assumption of proportional baseline risks as in [8]). In the special case of a leaky intervention, this is equivalent to a condition that we call "balanced replacement," which requires that for each subject, the exposure type be independent of the exposure time and exposure history. We show that if there is subject variation in infection risk, then even balanced replacement is insufficient to ensure the proportional pseudohazards condition.

A sieve effect is defined as any violation of equivalence of VE_s^{ϕ} across mark types s [6]. We define per-exposure mark-specific relative risks:

$$\begin{aligned}\mathrm{RR}^{\phi}(s) &\equiv \frac{\Pr(\text{fail with type } s \mid \text{one exposure to type } s, \text{ vaccine recipient})}{\Pr(\text{fail with type } s \mid \text{one exposure to type } s, \text{ placebo recipient})} \\ &= \frac{\phi_{vs}}{\phi_{ps}},\end{aligned} \tag{4}$$

and let $\mathrm{VE}_s^{\phi} = 1 - \mathrm{RR}^{\phi}(s)$.

Thus, a sieve effect is defined as a lack of equivalence across all types $s \in 1, \ldots, J$ of $\mathrm{RR}^{\phi}(s)$. In terms of odds ratios to some baseline type (arbitrarily we use type $s = 1$ here), the null hypothesis of no sieve effect is that for all s, $\mathrm{OR}^{\phi}(s) = 1$, where

$$\begin{aligned}\mathrm{OR}^{\phi}(s) &= \frac{\mathrm{RR}^{\phi}(s)}{\mathrm{RR}^{\phi}(1)} \\ &= \frac{\phi_{vs}/\phi_{ps}}{\phi_{v1}/\phi_{p1}} \\ &= \frac{\phi_{vs}/\phi_{v1}}{\phi_{ps}/\phi_{p1}}.\end{aligned} \tag{5}$$

In Appendix B we revisit the proof that under the condition that was called "Assumption 2" in [6, page 804], which "implies that the strain-specific exposure intensities are proportional, that is, $\lambda_{Es}(t) = \theta_s \lambda_{E1}(t)$," estimates of odds ratios based only on the type distributions of observed infections in treated and untreated subjects of a randomized controlled trial are unbiased for $\mathrm{OR}^{\phi}(s)$. The proof establishes an equivalence between the per-exposure odds ratio $\mathrm{OR}^{\phi}(s)$ and two other odds ratios of interest: the "prospective" (or "attack-rate") odds ratio, $\mathrm{OR}^{a}(s)$ and the "retrospective" odds ratio $\mathrm{OR}^{r}(s)$, as defined below. We show in Appendix A that the proof not only relies on the assumption, following [9], that for any subject the type-specific exposure hazard is that of a history-independent (zero-order) process, but also that it relies on the stronger assumption that the "exposure pseudohazards" are proportional. Whereas exposure hazards condition on the exposure processes, the pseudohazards condition on the subject's infection count being $N(t) = 0$, which depends both on the exposure processes E and on the chance of each exposure resulting in an infection.

In Appendix C we show that unless each subject can experience at most one exposure during the trial (a condition that we call "thoroughly rare events"), proportional pseudohazards require independence between each subject's exposure type distribution and the timing of his exposures (a condition that we call "balanced replacement"). As noted in [8], this in turn implies independence between each subject's infection time, T, and the mark of his infection, S, a condition that could only hold under a null hypothesis of no sieve effects. We then show that the proof requires subject homogeneity in risk. Risk inhomogeneity leads to a violation of the proportional pseudohazards condition, and of $T \perp\!\!\!\perp S$, even when the balanced replacement condition holds for each risk group (or individual subject) separately.

In Appendix D we revisit the argument that time-to-event methods such as competing risks Cox proportional hazards models can yield unbiased estimates of these quantities even when "Assumption 2" is violated. We argue that the assertion of unbiasedness requires an assumption of "noninformative censoring" when treating infections with some marks as censoring events while evaluating other mark types of infections. Since this implies that $T \perp\!\!\!\perp S$, we argue that the time-to-event methods are also biased unless Assumption

2 holds. We show that heterogeneity of the intervention effect across subjects will generally lead to a violation of the noninformative censoring assumption.

From these arguments we conclude that existing methods for evaluating hypotheses of sieve effects of leaky vaccines are expected to be biased if there is any subject heterogeneity in risk (or in response to the treatment), or if replacement failures are imbalanced. Gilbert has evaluated bias under violations of Assumption 2 in simulation studies [7, 8] and showed that under the conditions of those simulations the bias is limited to a few percentage points unless the marginal attack rate a_p is substantial, even when some subjects have no response to the intervention at all. However, those simulations ensured equal exposure distributions across subjects and did not carefully control replacement distribution balance. Future work is required to update the simulations to more specifically address issues of balanced replacement and of heterogeneous infecting exposure rates.

Particular caution is warranted when using case-only sieve analysis methods as introduced in [10], to which these arguments doubly apply, since in addition to the cautions expressed about the effects of risk group enrichment on case-only methods, the proof of the method's approximate unbiasedness depends on the assumption that individual mark-specific hazards can be evaluated by censoring other marks, using the noninformative censoring assumption. Since that assumption is surely violated whenever there are sieve effects, the use of the case-only sieve analysis method to evaluate sieve effects for subject-genotype dependency as the authors propose (or any other correlate) is not justified by the proof. Even under the null hypothesis of no sieve effects, the arguments presented here and in Appendix D show that the noninformative censoring assumption would only be reasonable in a setting in which types of distributions do not vary by risk group. If the different risk groups tend to be infected by different distributions of viruses even in the absence of treatment, as may be the case for HIV-1 trials (where risk is associated with mode of transmission, which in turn is associated with different populations of viruses), the assumption is likely violated.

It remains likely that these methods, though not proven unbiased, retain their power to detect sieve effects under the heterogeneous risk conditions that we have considered. Although the conditions of those proofs may not hold under heterogeneity, we have not proven the contrary assertion; other proofs that establish conditions under which unbiased estimation is robust to subject variation in risk may yet be devised. Also, in practice absolute unbiasedness may not be required; with further work evaluating the practical implications of these insights, we expect that these methods will be approximately unbiased for many or most applications to leaky vaccines with heterogeneous risk. It remains to future work to conduct a thorough evaluation of the loss of power or the potential anticonservatism of analyses that assume risk homogeneity when the assumption is not justified.

4. Conclusions

In this paper we have restated the argument that when conducting statistical analysis of vaccine efficacy trials with heterogeneous exposure or susceptibility risk, care should be taken to account for the putative mechanism of the vaccine. Two extremes of the spectrum of vaccine mechanisms are considered. At one extreme (all-or-none), a vaccine protects some fraction of subjects completely and the remaining fraction are unaffected by it. At the opposite extreme (leaky), a vaccine reduces the per-exposure transmission rate for all recipients equally. We have shown that leaky vaccines induce a violation of the proportional hazards condition that is often assumed in survival analysis, due to a changing fraction of at-risk subjects over time in both vaccinated and unvaccinated individuals. Since these fractions change over time differently in the two treatment groups, even if the proportional hazards condition holds for each risk group individually, the marginal hazard ratio changes over time.

Another effect of subject risk heterogeneity in leaky vaccine trials is that the relative proportions of the risk groups among infected subjects changes over time. We showed that associations between subject covariates and vaccine efficacy will be biased unless those covariates are distributed equivalently in all risk groups. A simple diagnostic analysis of the risk of infection among placebo recipients as a function of the covariate could be used to reject the hypothesis of independence that is required for interpreting correlations with vaccine efficacy as indicative of differential efficacy rather than differential baseline risk, but this is not possible in a "case-only" analysis (which evaluates the association only among infected subjects). This argument cautions against case-only analysis of correlates of the partial efficacy of leaky vaccines when there is subject heterogeneity in risk.

We also addressed the context of competing risks and showed that leaky vaccines with risk heterogeneity will induce time variation in the relative proportion of marks (types of the competing risks) of infections and that since this time variation occurs at different rates in the vaccine and placebo groups, this induces a violation of the equivalence between observable relative attack rates and unobservable per-exposure relative risks that is required for unbiased analysis of mark-specific vaccine efficacies (called "sieve effects" when they differ across types) [6]. Furthermore, this scenario has implications for the commonly encountered analysis methodology of analyzing one mark type of the competing risks by treating the infections by any other type as right-censoring events. In particular, the censorship process will not be independent of the infection process unless the infection times of the competing risks are independent, but the changing fractions of risk groups among the at-risk subjects induce dependence (even when the processes are conditionally independent).

Longini and Halloran [11] introduced an approach (frailty models) to evaluating a vaccine's efficacy when subject susceptibilities in any treatment group vary (with some fraction experiencing complete immunity as in an all-or-none vaccine and the remainder having some per-exposure susceptibility that may vary across individuals and differently

for each treatment group). This approach enables estimation of more complex vaccine effects, but the observations in this paper about the implications of leaky vaccines with subject heterogeneity in risk apply to these mixture models too, so for instance a relative enrichment of high-risk subjects among infected vaccine recipients cautions against naive correlation of risk-dependent covariates with infection outcomes.

Recent work has introduced sieve analysis methods for nonleaky vaccines, which have all-or-none style protection but perhaps against only a subset of risk mark categories (in which case they are called "some-or-none" vaccines) [12]. The all-or-none and some-or-none scenarios also engender differential enrichment of high-risk subjects among infected (and also at-risk) subjects, but because any protected subject is fully protected against the vaccine-targeted mark types, the attack rate will be reduced equally across risk groups as long as risk is independent of relative exposure rates. If, however, risk (in terms of rates of infection) is associated with the mark among placebo recipients, as expected for example, in HIV-1 vaccine trials, then the vaccine will reduce infection rates for one risk group more than another.

The arguments in this paper together imply that it is generally not possible to differentiate between mark-specific efficacy and subject-covariate-specific efficacy using failure time and failure type data alone unless subject risk is homogeneous. The only exception is when risk groups (though heterogeneous in overall failure rate) have homogeneous relative rates of the marks of infecting exposures across competing risk mark types. Future work is needed to develop statistical analysis methods that account for both subject heterogeneity (as in a frailty model) and competing risks such that the effects of each can be differentiated in an analysis of a partially efficacious vaccine. Such approaches would likely require parameterization not just of the frailty model but also of the exposure processes, requiring considerable modeling effort and sensitivity analysis.

Appendices

A. Correction to the Definition of Proportional Exposure Pseudohazards

In practice we are usually unable to observe exposure events (as noted in [9]), complicating estimation of the per-exposure probabilities of failure ϕ_{xs}. We observe the "retrospective" mark type distributions P^r_{xs} among those who become infected before the end of trial:

$$P^r_{xs} \equiv \Pr\left(\text{infected with mark } s \mid \text{infected in } [0,\tau],\ \text{treatment assignment is } x\right). \tag{A.1}$$

This is distinct from the "prospective" (or "joint attack rate") mark type distribution P^a_{xs}, which is the joint probability of infection ("failure") (occurring at all) and that the failure is of type s. It can also be defined in terms of the "retrospective" failure type distribution P^r_{xs} and the marginal failure probability a_x, since

$$\begin{aligned} a_x &\equiv \Pr(\text{failed in } [0,\tau] \mid x), \\ P^a_{xs} &\equiv \Pr\left(\text{fail with type } s \text{ in } [0,\tau] \mid x\right) \\ &= a_x \times P^r_{as}. \end{aligned} \tag{A.2}$$

Both the retrospective and prospective probabilities are distinct from the per-exposure probabilities that we intend to estimate. Gilbert et al. showed in [6] that, under certain conditions, there is an equivalence between the odds ratios for all three of these. The proof (repeated below in Appendix B) begins with an expression of the failure hazard as a function of the type-specific rate of exposure $\lambda_{Es}(t)$ and the per-exposure probabilities of failure ϕ_{xs} [6, page 805]. With T defined as the time of the subject's first failure, S the mark of that failure, and τ defined as the duration of the clinical trial, he wrote that, for any $t \in [0,\tau]$,

$$\begin{aligned} &\Pr(t \le T < t+\Delta t, S = s \mid T \ge t, x) \\ &\quad = \Pr\left(\text{exposed to type } s \text{ in } [t, t+\Delta t) \mid x, \text{ exposure history}\right) \\ &\quad \times \Pr(t \le T < t+\Delta t, S = s \mid T \ge t, x, \text{ exposed to } s \text{ in } [t, t+\Delta t), \text{ exposure history}). \end{aligned} \tag{A.3}$$

We note that the first term on the right hand side should include the condition $T \ge t$ (that the subject has not yet experienced a failure as of time t). This is not subsumed by "exposure history" because it depends not only on the exposure process(es) but also on the per-exposure probabilities of failure. Also, the "exposure history" condition does not exist on the left-hand side of the equation, so it should either be added there or removed from the right-hand side. That is, we can define the exposure-history conditional failure hazard as

$$\begin{aligned} &\Pr\left(t \le T < t+\Delta t, S = s \mid T \ge t, x, \text{exposure history}\right) \\ &\quad = \Pr\left(\text{exposed to type } s \text{ in } [t, t+\Delta t) \mid T \ge t, x, \text{ exposure history}\right) \\ &\quad \times \Pr(t \le T < t+\Delta t, X = s \mid T \ge t, x, 1\text{in exposed to } s \text{ in } [t, t+\Delta t), \text{ exposure history}), \end{aligned} \tag{A.4}$$

and then define the *marginal* failure hazard $\lambda_{xs}(t)$ in terms of this as

$$\begin{aligned} &\sum_{\text{exposure history}} \Pr(t \le T < t+\Delta t, S = s \mid T \ge t, x, \text{ exposure history}) \\ &\quad \times \Pr\left(\text{exposure history} \mid T \ge t, x\right). \end{aligned} \tag{A.5}$$

Since the per-exposure failure probability is assumed to be independent of exposure history, we can directly define the marginal failure hazard by dropping the condition from the right-hand side:

$$\begin{aligned}\lambda_{xs}(t) &\equiv \Pr(t \le T < t+\Delta t, S = s \mid T \ge t, x)\\ &= \Pr(\text{exposed to type } s \text{ in } [t, t+\Delta t) \mid T \ge t, x)\\ &\quad \times \Pr(t \le T < t+\Delta t, S = s \mid T \ge t, x,\\ &\qquad \text{exposed to } s \text{ in } [t, t+\Delta t)).\end{aligned} \tag{A.6}$$

We define the first term of this corrected failure hazard as the "exposure pseudohazard":

$$\ell_{xs}(t) = \Pr(\text{exposed to type } s \text{ in } [t, t+\Delta t) \mid T \ge t, x). \tag{A.7}$$

This is distinct from the generalized hazard function of the type s exposure process E_s:

$$\Pr(\text{exposed to type } s \text{ in } [t, t+\Delta t) \mid x, \text{exposure history}), \tag{A.8}$$

which conditions only on its own history (a generalization of the standard hazard function's dependence on nonfailure to time t). There is no difference if every exposure results in a failure, but the two functions depart whenever any exposure event could be avoided (by a roll of the "leaky" dice, with probability $1 - \Pr(t \le T < t+\Delta t, S = s \mid T \ge t, x$, exposed to s in $[t, t+\Delta t)) = 1 - \phi_{xs}$).

Conceptually, if exposures are occurring that do not result in failures (we call these "avoided failures"), then the subject may nevertheless fail, but later than she would have otherwise. Since lower per-exposure failure probabilities result in more avoided failures, the time-to-event distribution among those who fail in the treated group will be right-shifted compared to what it would have been in the untreated group. We note that it may not be right-shifted for a particular failure type, but aggregating over all types, lower per-exposure failure rates will result in later expected failure times. Mathematically, this dependence between the probability of nonfailure by time t (i.e., that $T \ge t$) and the rates of failure avoidance $(1-\phi_{xs})$ can be shown by expanding $\ell_{xs}(t)$ in the equation $\lambda_{xs}(t) = \ell_{xs}(t)\phi_{xs}$:

$$\begin{aligned}\ell_{xs}(t) = \sum_{\text{exposure history}} &\Pr(\text{exposed to } s \text{ in}\\ &[t, t+\Delta t) \mid T \ge t, x,\\ &\text{exposure history})\\ &\times \Pr(\text{exposure history} \mid T \ge t, x).\end{aligned} \tag{A.9}$$

By Bayes' theorem,

$$\begin{aligned}&\Pr(\text{exposure history} \mid T \ge t, x)\\ &\quad \propto \Pr(T \ge t \mid \text{exposure history}, x)\\ &\qquad \times \Pr(\text{exposure history} \mid x).\end{aligned} \tag{A.10}$$

These are not all equal, since if the exposure history includes k exposure times $e_1, \ldots, e_k$, then $\Pr(T \ge t \mid$ exposure history, $x)$ involves a product of the k chances that those failures were avoided and would be monotonically decreasing as the number of exposures increases (except in the boring case of a perfect intervention).

In Appendix B we repeat the proof from [6, 3.12] of the equivalence of the odds ratios, using these corrected definitions. The proof crucially depends on the assumption that the exposure pseudohazards $\ell_{xs}(t)$ are proportional across types. That is, it requires that there exist J constants θ_s such that $\forall s$, $\ell_{xs}(t) = \theta_s \ell_{x1}(t)$. By (A.9), this would necessitate setting $\Pr($exposed to s in $[t, t+\Delta t) \mid T \ge t, x,$ exposure history) to depend on exposure history in such a way that exactly counteracts the effect of variation in $\Pr($exposure history $\mid T \ge t, x)$. This is the condition that we call "balanced replacement" (so-called because it requires that the conditional distribution of failure types to be the same regardless of the number of avoided failures, so "replacement failures" have the same distribution as the failures that they replace through failure avoidance and subsequent reexposure). It is difficult to imagine how this perfect balance could be accomplished other than by assuming complete independence between the type of the exposure and both its timing and the history of the exposure processes (as in independent Poisson-distributed exposure processes for each mark). Effectively, therefore, balanced replacement implies that for each subject (given his treatment and in general his response to the treatment, as discussed below), the time and type of his failures are independent.

B. Proof of the Equivalence of Odds Ratios under Proportional Pseudohazards

Here we repeat the proof, given in [6], of the equivalence of the retrospective odds ratios and the per-exposure odds ratios. The proof begins by using the equation $\Pr(T \ge t \mid x) = e^{-\Lambda(t|x)}$ relating a survivor function to a cumulative hazard function to establish that, under conditions of proportional exposure pseudohazards,

$$\begin{aligned}\Pr(T \ge t \mid x) &= \exp\left(-\int_0^t \lambda(u \mid x)\, du\right)\\ &= \exp\left(-\int_0^t \sum_l \lambda(u, l \mid x)\, du\right)\\ &= \exp\left(-\int_0^t \sum_l \ell_{xl}(u)\, \phi_{xl}\, du\right).\end{aligned} \tag{B.1}$$

Then the prospective probabilities can be written in terms of the exposure pseudohazards: as

$$\begin{aligned}P_{xs}^a &= \int_0^\tau \lim_{\Delta t \searrow 0} \frac{\Pr(t \le T < t+\Delta t, S = s \mid T \ge t, x)}{\Delta t}\, dt\\ &= \int_0^\tau \lambda(t, s \mid x) \times \Pr(T \ge t \mid x)\, dt\end{aligned}$$

$$= \int_0^\tau \ell_{xs}(t)\,\phi_{xs}\ \Pr(T \geq t \mid x)\,dt$$

$$= \int_0^\tau \ell_{xs}(t)\,\phi_{xs} \exp\left(-\int_0^t \sum_l \ell_{xl}(u)\,\phi_{xl}\,du\right) dt. \tag{B.2}$$

Then if we define the integrated type s exposure pseudohazard $F_{xs}(t) \equiv \int_0^t \ell_{xs}(u)du$, we get

$$\begin{aligned}
P_{xs}^a &= \int_0^\tau \ell_{xs}(t)\,\phi_{xs} \exp\left(-\int_0^t \sum_l \ell_{xl}(u)\,\phi_{xl}\,du\right) dt \\
&= \theta_s\phi_{xs} \int_0^\tau \ell_{x1}(t) \exp\left(-\int_0^t \ell_{x1}(u)\,du \sum_l \theta_l\phi_{xl}\right) dt \\
&= \theta_s\phi_{xs} \int_0^\tau \ell_{x1}(t) \exp\left(-F_{x1}(t) \sum_l \theta_l\phi_{xl}\right) dt \\
&= \left| \frac{-1}{\sum_l \theta_l\phi_{xl}} \exp\left(-u \sum_l \theta_l\phi_{xl}\right)\right|_0^{F_{x1}(\tau)} \times \theta_s\phi_{xs} \\
&= \left(1 - \exp\left(-F_{x1}(\tau) \sum_l \theta_l\phi_{xl}\right)\right) \frac{\theta_s\phi_{xs}}{\sum_l \theta_l\phi_{xl}} \\
&= (1 - \Pr(T \geq \tau \mid x)) \frac{\theta_s\phi_{xs}}{\sum_l \theta_l\phi_{xl}} \\
&= a_x \frac{\theta_s\phi_{xs}}{\sum_l \theta_l\phi_{xl}},
\end{aligned} \tag{B.3}$$

which, since $P_{xs}^a = a_x \times P_{xs}^r$, implies that

$$P_{xs}^r = \frac{\theta_s\phi_{xs}}{\sum_l \theta_l\phi_{xl}}. \tag{B.4}$$

This proof relies on the proportional pseudohazards condition to enable the factorization that separates the integrated exposure pseudohazard $F_{x1}(\tau)$ for an arbitrary mark ($s = 1$) from the time-constant multiples θ_s for $s > 1$ such that $F_{xs}(\tau) = \theta_s F_{x1}(\tau)$.

Finally, this result guarantees equivalence of retrospective, prospective, and per-exposure odds ratios, since

$$\begin{aligned}
\mathrm{OR}^r(s) &\equiv \frac{P_{vs}^r / P_{ps}^r}{P_{v1}^r / P_{p1}^r} \\
&= \frac{P_{vs}^r / P_{v1}^r}{P_{ps}^r / P_{p1}^r} \\
&= \frac{\theta_s\phi_{vs} / \theta_1\phi_{v1}}{\theta_s\phi_{ps} / \theta_1\phi_{p1}} \\
&= \frac{\phi_{vs} / \phi_{v1}}{\phi_{ps} / \phi_{p1}} \\
&= \mathrm{OR}^\phi(s), \quad \text{QED.}
\end{aligned} \tag{B.5}$$

C. Conditions for Proportional Pseudohazards

The result of Appendix B is limited to conditions of proportional exposure pseudohazards, which requires balanced replacement. Technically, balanced replacement means that for each subject, given her response to the intervention, any variation in her probability of exposure to a potential failure of type s due to dependence on time or exposure history must exactly counterbalance the effects of unavoidable variation in the probability of exposure history over time (conditioned on survival up to that time), such that the relative rate of exposure to one type over another type remains constant. This is accomplished if we assume that the type of the exposure is completely independent of the failure time (and history), an assumption that has been discussed elsewhere [7, 13].

C.1. Thoroughly Rare Events. Proportional exposure pseudohazards could also be accomplished if we assume that each subject experiences at most one failure during $[0, \tau]$ (because in that case, every exposure is a "first exposure" and there are no replacement failures; put another way, in that case the probability of survival given exposure history is effectively independent of exposure history, so there is nothing to balance). It may be tempting to argue that in rare-event settings, this is a reasonable assumption. We note however that the requirement of no replacement failures is stronger than a typical "rare event" scenario, in which the rates a_x are small. The condition we require (thoroughly rare events) means that for all subjects, the probability of a second exposure is zero. While in very-rare event settings, violations of this condition may not lead to bias, it should be noted that the relevant determinant is not the rareness of the event in the total population but the rareness of the event in the subset of subjects who experience a failure during the trial. Since in general, subjects who experience more than one exposure have more chances of failure, in expectation, the probability of experiencing more than one exposure is greater among the subjects who fail than among the larger population. In other words, there is an enrichment of multiply exposed subjects among the subjects who experience a failure during the trial. This is restating our main finding that high-risk subjects are enriched among infected vaccinees.

C.2. Risk Homogeneity. Gilbert noted that conditions of the proof require a "leaky" vaccine in which all subjects experience the same intervention effect, and he explored violations of this assumption through simulations [8]. He showed that if some subjects do not have "take" of the intervention (i.e., if there is a chance that an intervention-receiving subject might nevertheless have no change in her per-exposure probabilities of infection), a bias is introduced. Here we show why the proof breaks down in the context of incomplete take. It is analogous to the setting of dichotomous risk, since for high-risk subjects the leaky vaccine's effect on attack-rate vaccine efficacy is reduced, and in some cases it becomes effectively nil.

For the example with two risk groups, the exposure pseudohazard for vaccine recipients becomes

$$\ell_{vs}(t) = \omega_{hv}(t)\,\ell_{vhs}(t) + \left(1 - \omega_{hv}(t)\right)\ell_{vls}(t), \qquad \text{(C.1)}$$

where ℓ_{vhs} and ℓ_{vls} are the type s exposure pseudohazards for high-risk and low-risk subjects, respectively, and

$$\omega_{hx}(t) \equiv \Pr\left(\text{high risk} \mid \text{infected with mark } s \text{ by time } t, x\right). \qquad \text{(C.2)}$$

For placebo recipients analogously

$$\ell_{ps}(t) = \omega_{hp}(t)\,\ell_{phs}(t) + \left(1 - \omega_{hp}(t)\right)\ell_{pls}(t). \qquad \text{(C.3)}$$

Since $\omega_{hp}(t)$ varies over time and differently from $\omega_{hv}(t)$, the exposure pseudohazards depend on time and risk group and do not satisfy the proportionality condition, even if the risk group specific-exposure pseudohazards do satisfy it.

D. Homogeneous Intervention Effects and Proportional Pseudohazards Are Both Required for Noninformative Censoring

Gilbert argued in [8] that some time-to-event methods are robust to violations of the condition that we call "proportional exposure pseudohazards." He argued that estimates (of retrospective odds ratios) using Cox proportional hazards models without proportional baseline risks could be used to estimate each per-exposure odds ratio in a separate model. So for instance, the per-exposure odds ratio for strain s versus strain 1 could be estimated by treating all other types of failure as censoring events. Since the methodology for this estimation requires an assumption of noninformative censoring, this argument breaks down unless the time-to-event distribution of type 1 and type s failures is independent of that of the other failure types. Since the condition must hold for all choices of s, it effectively requires independence between the time-to-event distribution T and the failure type distribution S. In the case of a leaky vaccine with homogeneous risk, this can be achieved under the conditions of "balanced replacement" (which requires that the exposure processes exhibit the same sort of time/type independence that is desired for the failure hazard).

We now show that any amount of subject heterogeneity in the intervention effect will lead to a violation of the noninformative censoring assumption, except under the null hypothesis of no sieve effect. We have shown that infection time T is not independent from risk group R, and so if mark type S is also nonindependent from risk, then S and T will not be independent. If time and type are not independent then if you treat some marks of infection as censoring events, then those events are not independent of the uncensored infection times, and noninformative censoring does not hold.

So it remains to show that failure mark type S is not independent of risk. It is clearly the case that if high-risk subjects have a different mark distribution of exposures, then by assumption S depends on risk. Also, by the same mechanism of exposure-rate-dependent attack-rate efficacy that is discussed in Section 2, even if high-risk placebo recipients have the same mark distribution of exposures as low-risk placebo recipients, the mark-specific attack-rate vaccine efficacy will vary across the types unless the infecting exposure rates are identical across all types. If the mark distribution of exposures is same for both risk groups and the rate of exposures is constant across marks within each risk group, then only the overall rate of infection varies across risk groups, and the VE_s^a is the same for all marks. This condition clearly precludes sieve effects.

The noninformative censoring assumption will never hold if there are sieve effects, but even under the null hypothesis of no sieve effects the assumption will only hold in the extreme case in which the (placebo-recipient) infecting exposure processes for all marks are equidistributed within each risk group. As long as some mark exposures occur at higher rates than others, then the attack-rate vaccine effect will differ against the different types, leading to a violation of the noninformative censoring condition.

Conflict of Interests

The author declares that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The author thanks Betz Halloran, James Dai, Peter Gilbert, and Jason Shao for inspiration and feedback. The author retains all blame for errors or inaccuracies. This research was supported by NIH NIAID Grant 2 R37 AI054165-11. The opinions expressed in this paper are those of the author and do not represent the official views of the NIAID.

References

[1] P. H. Rhodes, M. E. Halloran, and I. M. Longini Jr., "Counting process models for infectious disease data: distinguishing exposure to infection from susceptibility," *Journal of the Royal Statistical Society B*, vol. 58, no. 4, pp. 751–762, 1996.

[2] M. E. Halloran, I. M. Longini, and C. J. Struchiner, *Designand Analysis of Vaccine Studies*, Springer, New York, NY, USA, 2010.

[3] M. E. Halloran, M. Haber, and I. M. Longini Jr., "Interpretation and estimation of vaccine efficacy under heterogeneity," *American Journal of Epidemiology*, vol. 136, no. 3, pp. 328–343, 1992.

[4] S. A. Plotkin and P. B. Gilbert, "Nomenclature for immune correlates of protection after vaccination," *Clinical Infectious Diseases*, vol. 54, no. 11, pp. 1615–1617, 2012.

[5] P. T. Edlefsen, P. B. Gilbert, and M. Rolland, "Sieve analysis in hiv-1 vaccine effcacy trials," Current Opinion in HIV and AIDS, 8:000-000, 2013.

[6] P. B. Gilbert, S. G. Self, and M. A. Ashby, "Statistical methods for assessing differential vaccine protection against human immunodeficiency virus types," *Biometrics*, vol. 54, no. 3, pp. 799–814, 1998.

[7] P. B. Gilbert, "Interpretability and robustness of sieve analysis models for assessing hiv strain variations in vaccine effcacy," *Statistics in Medicine*, vol. 20, no. 2, pp. 263–279, 2001.

[8] P. B. Gilbert, "Comparison of competing risks failure time methods and time-independent methods for assessing strain variations in vaccine protection," *Statistics in Medicine*, vol. 19, no. 22, pp. 3065–3086, 2000.

[9] M. E. Halloran, C. J. Struchiner, and I. M. Longini Jr., "Study designs for evaluating different efficacy and effectiveness aspects of vaccines," *American Journal of Epidemiology*, vol. 146, no. 10, pp. 789–803, 1997.

[10] J. Y. Dai, S. S. Li, and P. B. Gilbert, "Case-only method for cause-specific hazards models with application to assessing differential vaccine efficacy by viral and host genetics," *Biostatistics*, vol. 15, no. 1, pp. 196–203, 2014.

[11] I. M. Longini Jr. and M. E. Halloran, "A frailty mixture model for estimating vaccine efficacy," *Journal of the Royal Statistical Society. Series C*, vol. 45, no. 2, pp. 165–173, 1996.

[12] P. T. Edlefsen, "Evaluating the dependence of a non-leaky intervention's partial efficacy on a categorical mark," http://arxiv.org/abs/1206.6701.

[13] M. Juraska and P. B. Gilbert, "Mark-specific hazard ratio model with multi- variate continuous marks: an application to vaccine effcacy," *Biometrics*, vol. 69, no. 2, pp. 328–337, 2013.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 160520, 14 pages
http://dx.doi.org/10.1155/2014/160520

Research Article

Structural Equation Modeling for Analyzing Erythrocyte Fatty Acids in Framingham

James V. Pottala,[1,2] Gemechis D. Djira,[3] Mark A. Espeland,[4] Jun Ye,[5] Martin G. Larson,[6,7,8] and William S. Harris[1,2,9]

[1] *Health Diagnostic Laboratory Inc., Richmond, VA 23219, USA*
[2] *Department of Internal Medicine, Sanford School of Medicine, University of South Dakota, Sioux Falls, SD 57105, USA*
[3] *Department of Mathematics and Statistics, South Dakota State University, Brookings, SD 57007, USA*
[4] *Department of Biostatistical Sciences, Wake Forest School of Medicine, Winston-Salem, NC 27157, USA*
[5] *Department of Statistics, University of Akron, Akron, OH 44325, USA*
[6] *Department of Biostatistics, Boston University School of Public Health, Boston, MA 02218, USA*
[7] *Department of Mathematics and Statistics, Boston University, Boston, MA 02215, USA*
[8] *Framingham Heart Study, Framingham, MA 01702, USA*
[9] *OmegaQuant Analytics, Sioux Falls, SD 57107, USA*

Correspondence should be addressed to James V. Pottala; jpottala@hdlabinc.com

Received 26 December 2013; Revised 28 February 2014; Accepted 28 February 2014; Published 15 April 2014

Academic Editor: Zhenyu Jia

Research has shown that several types of erythrocyte fatty acids (i.e., omega-3, omega-6, and *trans*) are associated with risk for cardiovascular diseases. However, there are complex metabolic and dietary relations among fatty acids, which induce correlations that are typically ignored when using them as risk predictors. A latent variable approach could summarize these complex relations into a few latent variable scores for use in statistical models. Twenty-two red blood cell (RBC) fatty acids were measured in Framingham ($N = 3196$). The correlation matrix of the fatty acids was modeled using structural equation modeling; the model was tested for goodness-of-fit and gender invariance. Thirteen fatty acids were summarized by three latent variables, and gender invariance was rejected so separate models were developed for men and women. A score was developed for the polyunsaturated fatty acid (PUFA) latent variable, which explained about 30% of the variance in the data. The PUFA score included loadings in opposing directions among three omega-3 and three omega-6 fatty acids, and incorporated the biosynthetic and dietary relations among them. Whether the PUFA factor score can improve the performance of risk prediction in cardiovascular diseases remains to be tested.

1. Introduction

Higher blood levels of the essential omega-3 polyunsaturated fatty acids (PUFA) are associated with reduced risk for sudden cardiac death [1, 2] and all-cause mortality [2, 3]. There is also evidence that the essential omega-6 PUFA intakes and blood levels are inversely associated with risk for coronary heart disease [4]. Other fatty acids, such as the *trans* fatty acids found in partially hydrogenated vegetable oils, are believed to increase risk for cardiovascular disease [5]. Hence, the study of these fatty acids is of vital importance.

PUFA are "essential" since they cannot be produced *in vivo* and must be consumed. Foods are composed of multiple fatty acids, and dietary habits manifest themselves as correlated fatty acid levels in the blood. Once consumed, simpler PUFA species can be acted upon by enzymes that convert them into more complex PUFA which have a wide variety of metabolic functions. Desaturase enzymes insert double bonds (points of "desaturation") into fatty acid molecules, and elongase enzymes are needed to increase the carbon chain length [6]. Importantly, competition among PUFA species exists for these enzymes such that different ratios in the diet can affect overall PUFA patterns (Figure 1) [7]. While elongase enzymes are readily available, the desaturase enzymes are rate limiting, and thus their levels may impact the amount of the 20- and 22-carbon fatty acids present

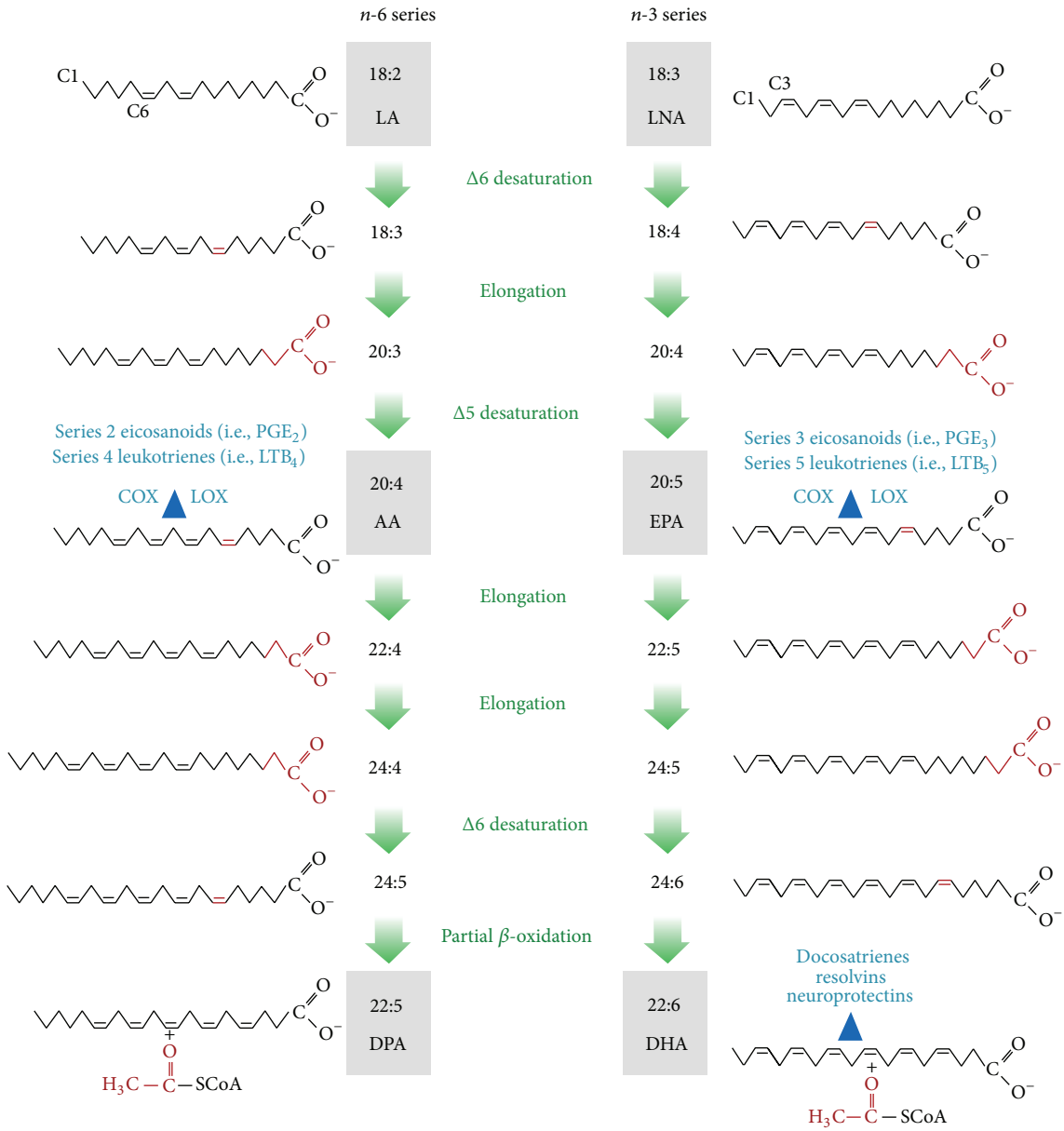


Figure 1: Polyunsaturated fatty acid biosynthesis [7]. Permission to reproduce this figure was granted on December 26, 2013, from the journals' copyright clearance center. Biosynthesis of long-chain *n-3* and *n-6* series polyunsaturated FAs from their 18-carbon precursors. The terminal methyl group is carbon 1 and the *n-3* and *n-6* series of FAs are termed according to the position of the first double bond: after carbon 3 and carbon 6, respectively. Biologically important FAs are highlighted with a gray box. Newly added/removed carbons or double bonds introduced at each step are colored red. Signaling molecules derived from AA, EPA, and DHA are noted in blue. LA, linoleic acid; LNA, linolenic acid; AA, arachidonic acid; EPA, eicosapentanoic acid; DPA, docosapentanoic acid; DHA, docosahexaenoic acid; COX, cyclooxygenase; LOX, lipoxygenase; PG, prostaglandin; and LT, leukotriene.

in the system [8]. Additionally, omega-3 fatty acids are the preferential substrates over omega-6 for these desaturase enzymes [6]. These biochemical and dietary relations induce a correlation structure in the blood fatty acids.

Fatty acids have been reported as weight%, mol%, or concentration (by volume or cell count). Since there are no laboratory standards in the USA to uniformly report fatty acid data, these multiple presentations exist. The main debate is relative versus absolute amounts, whose quantities become more divergent as they increase [9]. Chow argues in favor of absolute concentrations, citing the obvious drawback of relative amounts being the imposed linear constraint (i.e, summation to 100%) [10]. However, Crowe prefers relative amounts since absolute concentrations rise and fall with total cholesterol, which is made up of lipoproteins composed of fatty acids [11]. Fatty acid nomenclature is as follows: C#:#n# = the number of carbon (C) atoms in the molecule, the number of double bonds, and the omega family (n), whether 3, 6, 7, or 9. The latter indicates on which carbon the final double bond resides. In Bradbury et al. concentrations and mol% are

compared for C14:0 and C18:2n6 in plasma cholesterol ester and phospholipids [12]. The study results show that C18:2n6 is significantly *directly* correlated with total cholesterol when represented as a concentration (μmol/L) but significantly *inversely* related as mol%. The paper lists several references that support the cholesterol lowering effect of C18:2n6 and concludes that the metabolic pathways are influenced by the percentage of total fatty acids and not by concentration. The other advantage of weight% representation is that RBC fatty acids have a strong correlation with myocardium tissue fatty acids ($r = 0.82$) [13] and dietary intake [14]. This preferred technique of using relative weight% of total fatty acids also induces a correlation structure in the data.

Structural equation modeling (SEM) is well suited to incorporate the metabolic, dietary, and measurement correlations observed in fatty acid data. Only in the last decade has SEM been applied to fatty acids [15–18]. SEM allows complex high-dimensional relations to be simplified into a few latent variable scores, which can be evaluated as novel risk markers. Sex-specific risk prediction models have been implemented for coronary heart disease [19] to account for gender differences in the amount of risk attributable to cholesterol and blood pressure levels. Similarly the observed fatty acids may relate to the underlying latent constructs differentially for men and women, and this potential gender invariance needs to be evaluated. A recent taxonomy has been developed to specifically test measurement invariance over multiple groups in SEM by comparing models with different constraints applied to the correlation structure [20, 21].

The objective of the present study was to reduce the dimensionality of the complex fatty acid correlation structure by incorporating dietary intake patterns and biosynthesis processes as constraints in a structural equation model. This technique was applied to the Framingham Offspring/Omni RBC samples, and differences in fatty acid means, loadings, residuals, and latent variable covariance structures between men and women were tested.

2. Materials and Methods

2.1. Materials. The Framingham Heart Study (FHS) was established in 1948 to research the factors that contribute to cardiovascular disease. Its study design and methods are described at http://www.nhlbi.nih.gov/about/framingham. In 1971, the children (and their spouses) of the original FHS were recruited; they constitute the Framingham Offspring cohort [22]. In 1994, to better reflect the changing demographics of the area, recruitment began for Framingham residents aged 40–74 who described themselves as members of a minority group, that is, Omni cohort [23]. The Offspring and Omni cohorts were scheduled together for comprehensive examinations every 4–8 years. These included anthropometric measurements, biochemical assessment for CVD risk factors, medical history, and physical examination by a study physician. RBC samples taken from Offspring Examination 8 and Omni Exam 3 (2005–2007) were collected and subsequently 22 fatty acids were analyzed using gas chromatography (GC), and their content was expressed as a weight% of total fatty acids [24]. These data are publicly available as part of the National Heart Lung and Blood Institute (NHLBI) SNP Health Association Resource (SHARe) project (release date: March 26, 2013, dataset name: l_rbcfa_2008_m_0420s, dataset accession: pht002568). Written informed consent was provided by all participants, and the Institutional Review Board at the Boston University Medical Center approved the study protocol.

The study participants had a mean age (SD) of 66 (9) years, 55% were female, and 91% were white (see Table 1 in Supplementary Material available online at http://dx.doi.org/10.1155/2014/160520). The prevalence of chronic disease was diabetes (14%), heart disease (10%), and congestive heart failure (2%). The participants were taking hypertension medications (49%), lipid pharmacotherapy (43%), aspirin 3+ per week (43%), and fish oil supplements (10%).

2.2. Methods

2.2.1. Model Notation. SEM is a two-part modeling process. The first part defines a measurement model, which specifies the relations between the fatty acids and the latent variables, given by $\mathbf{y}_i = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i$ $(i = 1, \ldots, N)$ [25] where $\mathbf{y}_i$ is a $p \times 1$ vector of observed fatty acids measured on subject i, $\boldsymbol{\nu}$ is a vector of fatty acid means, $\boldsymbol{\Lambda}$ is a matrix of unknown loading parameters, $\boldsymbol{\eta}_i$ is a $m \times 1$ vector of latent variable scores for subject i, and $\boldsymbol{\varepsilon}_i$ is vector of normal random errors with covariance matrix $\boldsymbol{\Theta}$ that is independent of $\boldsymbol{\eta}_i$. The second part defines a path model for the latent variables $\boldsymbol{\eta}_i$, which allows regressing one latent variable $\boldsymbol{\eta}_{1i}$ on the set of other latent variables $\boldsymbol{\eta}_{2i}$, given by $\boldsymbol{\eta}_{1i} = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta}_{2i} + \boldsymbol{\xi}_i$, where $\boldsymbol{\alpha}$ is a vector of latent variable means, $\mathbf{B}$ is a vector of unknown regression parameters, and $\boldsymbol{\xi}_i$ is vector of normal random errors with covariance matrix $\boldsymbol{\Psi}$.

2.2.2. Comparing SEM to Other Multivariate Techniques. The proportion of variance explained in the jth fatty acid by the latent variables is defined as the jth communality. The variance in each fatty acid is equal to its communality plus its unique variance; that is, $\sigma_{jj} = \lambda_{j1}^2 + \lambda_{j2}^2 + \cdots + \lambda_{jm}^2 + \theta_j$. Principal components analysis (PCA) ignores the specific variance (measurement error) and uses the identity matrix for θ_j. This results in factoring the *total* variance instead of the *common* variance; the latter is the proportion of variance shared by the fatty acids. If communalities < 1 are used with principal components method of decomposing the observed correlation matrix using eigenvalues and eigenvectors, then the method is principal factoring. This is akin to using the reduced sample correlation matrix where the main diagonals are less than one. Likewise, an exploratory factor analysis imposes no structure and also assumes independence in the residuals matrix, but the common variance is extracted and the factors can be correlated through oblique rotations. Moving to confirmatory factor analysis requires restrictions on the model parameter, which allows testing the model goodness-of-fit. However, only structural equation modeling (SEM) allows regression paths among the observed and

Table 1: Framingham subjects' fatty acids; mean (SD).

Fatty acid	Overall $N = 3196$	Male $N = 1434$	Female $N = 1762$	P value*
Myristic, C14:0	0.31 (0.08)	0.29 (0.07)	0.32 (0.09)	<**0.0001**
Palmitic, C16:0	21.29 (1.24)	21.24 (1.21)	21.34 (1.26)	0.031
Stearic, C18:0	18.11 (0.95)	18.20 (0.89)	18.04 (1.00)	<**0.0001**
Lignoceric, C24:0	0.43 (0.16)	0.44 (0.16)	0.42 (0.16)	0.012
Palmitoleic, C16:1	0.35 (0.19)	0.31 (0.18)	0.39 (0.19)	<**0.0001**
Oleic, C18:1	13.88 (1.06)	13.90 (1.06)	13.85 (1.06)	0.23
Eicosenoic, C20:1	0.27 (0.11)	0.28 (0.12)	0.27 (0.10)	<**0.0001**
Nervonic, C24:1	0.45 (0.15)	0.46 (0.15)	0.43 (0.15)	<**0.0001**
trans Palmitoleic, C16:1 trans	0.17 (0.05)	0.16 (0.05)	0.17 (0.05)	0.0052
trans Oleic, C18:1 trans	1.62 (0.55)	1.62 (0.57)	1.61 (0.54)	0.56
trans Linoleic, C18:2 trans	0.25 (0.08)	0.24 (0.08)	0.25 (0.08)	**0.0021**
alpha-Linolenic, C18:3n3	0.19 (0.10)†	0.17 (0.09)	0.20 (0.11)	<**0.0001**
Eicosapentaenoic (EPA), C20:5n3	0.74 (0.46)†	0.71 (0.42)	0.76 (0.49)	**0.0011**
Docosapentaenoic, C22:5n3	2.74 (0.46)	2.80 (0.46)	2.70 (0.45)	<**0.0001**
Docosahexaenoic (DHA), C22:6n3	4.88 (1.38)	4.82 (1.39)	4.92 (1.37)	0.043
Linoleic, C18:2n6	11.19 (1.74)	11.03 (1.63)	11.33 (1.81)	<**0.0001**
gamma-Linolenic, C18:3n6	0.08 (0.09)	0.08 (0.12)	0.09 (0.07)	0.089
Eicosadienoic, C20:2n6	0.28 (0.05)	0.28 (0.05)	0.28 (0.05)	0.23
Eicosatrienoic, C20:3n6	1.59 (0.36)	1.59 (0.36)	1.59 (0.35)	0.66
Arachidonic, C20:4n6	16.78 (1.62)	16.79 (1.57)	16.77 (1.66)	0.66
Docosatetraenoic, C22:4n6	3.76 (0.83)	3.92 (0.83)	3.63 (0.81)	<**0.0001**
Docosapentaenoic, C22:5n6	0.66 (0.19)	0.67 (0.19)	0.64 (0.19)	<**0.0001**

*Two-sample t-test, the critical level alpha was set to 0.05/22 = 0.0023 for statistical significance using Bonferroni correction (shown in bold). †Use the following mean (SD) of the log-transformed values when standardizing data as explained in the discussion section for C18:3n3 −6.38 (0.41) and C20:5n3 −5.04 (0.48).

unobserved variables, also a sparse correlation matrix for the residuals can be specified. Therefore, SEM allows the most model flexibility for implementing the dietary intake patterns and metabolic processes among the fatty acids.

2.2.3. Data Preparation. SEM requires multivariate normality for maximum likelihood (ML) estimation. We started by assessing univariate normality which is implied by multivariate normality. To examine univariate normality, skewness and kurtosis measures were calculated for each fatty acid. If a fatty acid had an absolute kurtosis >10, considered problematic [26], or skewness >3, then a natural logarithm transformation was employed. Next, the null hypothesis of multivariate normality was tested using the SAS macro %MULTNORM which calculates the squared Mahalanobis distances (D^2); for large samples, D^2 is distributed as a χ^2_p [27]. When multivariate normality is not tenable, robust ML estimation should be implemented [28].

The scales of the RBC fatty acids differed by two orders of magnitude, on average Palmitic acid (C16:0) accounted for 20% of total fatty acid abundance, whereas alpha-linolenic acid (C18:3n3) accounted for only 0.2%. However, there are meaningful fatty acids even at small relative weight%. For example, the mean levels of C20:5n3 and C16:1t are 0.7% and 0.2%, but these are widely studied biomarkers of fish and dairy intake, respectively. Therefore, all fatty acids were standardized in order to have similar effect sizes, which prevented the large abundance fatty acids from dominating the variance extraction. The measure of sampling adequacy developed by Kaiser [29] MSA = $\Sigma(\text{simple correlations})^2/[\Sigma(\text{simple correlations})^2 + \Sigma(\text{partial correlations})^2]$ was used to identify fatty acids that were not sufficiently related to the core latent structure. Individual fatty acids with a MSA < 0.60 were considered "unacceptable" [30] and dropped from analysis.

2.2.4. Gender Invariance Testing. The primary motivation of the study was to reduce the dimensions of the fatty acid correlation matrix and to develop latent variable scores for each subject using the regression scoring method [31]. An assumption when using latent variable scores is that the indicators (i.e, fatty acids) have the same relations with the underlying latent variables between groups of interest, in this case gender. To this end, the correlation matrix was partitioned as $\boldsymbol{\Sigma} = \boldsymbol{\Lambda\Psi\Lambda}^{\mathbf{T}} + \boldsymbol{\Theta}$, and it along with the mean profile $\boldsymbol{\nu}$ of the observed fatty acids was tested for gender invariance using the likelihood ratio test (LRT) by comparing models with different imposed constraints. When using robust maximum likelihood the LRT has been modified

by the deviance scaling [32]. The specific hypotheses are as follows.

(1) The fatty acid means are equal between genders, $H_1 : \boldsymbol{\nu}_{male}=\boldsymbol{\nu}_{female}$.
(2) The loadings matrix is equal between genders, $H_2 : \boldsymbol{\Lambda}_{male}=\boldsymbol{\Lambda}_{female}$.
(3) The fatty acid variances are equal between genders, $H_3 : \boldsymbol{\Theta}_{male}=\boldsymbol{\Theta}_{female}$.
(4) The latent variable covariances are equal between genders, $H_4 : \boldsymbol{\Psi}_{male}=\boldsymbol{\Psi}_{female}$.
(5) The fatty acid covariances are equal between genders, $H_5 : \boldsymbol{\Sigma}_{male}=\boldsymbol{\Sigma}_{female}$.

The number of underlying dimensions was examined using exploratory factory analysis where eigenvalues >1 were retained. Then a SEM was built using the same number of latent variables, and it was used to test if the fatty acid means $\boldsymbol{\nu}$ were equal between men and women (H_1). This was done by testing the model X^2 between a model with intercepts, loadings, and unique variances freely estimated for men and women (model M0) versus one with a single set of intercepts imposed for both genders (model M1). To test for equality in the loading matrix $\boldsymbol{\Lambda}$ between genders (H_2), a model with equal loading constraints, but freely estimated intercepts and unique variances for each gender (model M2), was compared to model M0. Hypothesis H_3 was tested for gender differences in fatty acid residual variances $\boldsymbol{\Theta}$ by comparing a model with freely estimated intercepts and loadings for each gender, but with constrained variances (Model M3) versus model M0. To test the latent variable covariance structure $\boldsymbol{\Psi}$, model M2 was used for comparison with additional constraints placed on the six latent covariances to be equal between genders (Model M4). Lastly the fatty acid covariance matrix was tested by constraining loadings, latent covariances, and unique variances to be equal for men and women (Model M5). Each model used the direct Quartimin oblique rotation (available in Mplus and SAS).

2.2.5. Comparing Model Fits. To evaluate the fit of the SEM there are several indices, and the following is the minimal set established by current practice: (1) model chi-square, (2) Steiger-Lind root mean square error of approximation (RMSEA), (3) Bentler comparative fix index (CFI), and (4) standardized root mean square residual (SRMSR) [26]. The RMSEA indicates the discrepancy in model fit per degree of freedom as defined by $\varepsilon = \text{sqrt}[(\chi^2 - \text{df})/(\text{df} \times (\text{N} - 1))]$ [33]. The RMSEA follows a noncentral X^2 distribution, which allows reversing the role of the null hypothesis to testing a poorly fitting model and then a larger sample provides evidence of good fit. RMSEA is not used to test for perfect fit $\varepsilon = 0$ but to test the alternative hypothesis of "close fit" $H_a : \varepsilon \leq 0.05$ or "reasonable fit" $H_a : \varepsilon \leq 0.08$ [26]. Bentler's CFI and the Tucker-Lewis Index (TLI) are relative fit indexes; these measures should be >0.90 [34], and CFI differences of 0.01 between models are considered relevant [35]. The absolute model fit was assessed by calculating the SRMSR between the fatty acids' observed correlations and the correlations predicted by the latent variables; these residuals should be less than 0.10 for a good fitting model [26]. The Schwarz Bayesian Criterion [36], which includes a larger penalty for lack of parsimony than Akaike Information Criteria [37], was also reported. Analyses were performed using SAS software (version 9.2; SAS Institute Inc., Cary, NC) and Mplus (version 6.12; Muthen & Muthen, Los Angeles, CA).

3. Results

3.1. Exploratory Factor Analysis. Table 1 indicates gender differences in mean levels, in 12 out of 22 RBC fatty acids. The greatest relative differences were higher levels of C16:1 and C18:3n3 in females. The largest absolute differences were that females had about 0.3 percentage point higher and lower levels of C18:2n6 and C22:4n6 than males, respectively. Skewness and kurtosis were calculated for the individual fatty acids, and the following had distributions with an absolute kurtosis index >10 and/or a skew index >3, that is, C20:1, C18:3n3, C20:5n3, and C18:3n6, which became approximately Gaussian using a natural logarithm transformation. However, about 60% of C18:3n6 measurements were <0.1%, which is considered as the reliable detection limit for the GC method, and it appears that the log transformation simply produced normally distributed noise. Therefore, C18:3n6 was excluded from latent variable analysis. Even though univariate normality was reasonable for the individual fatty acids, multivariate normality was rejected ($P < 0.0001$) so robust ML method was implemented in Mplus [28].

Fatty acid concentrations were standardized to produce a correlation matrix (Supplemental Table 2), and the MSA was calculated for each fatty acid and overall (the latter was initially 0.20). The fatty acid with the lowest MSA value was dropped from analysis until all fatty acid MSA values were >0.60. This resulted in the following fatty acids being sequentially excluded from the latent variable analysis: C18:1, C20:1, C18:2n6, C20:2n6, C20:3n6, C24:0, and C24:1. After excluding these fatty acids the overall MSA for the correlation matrix increased to 0.75. Additionally, C22:5n3 needed to be removed from the correlation matrix because it was causing the explained variability in C20:5n3 to be greater than 100%, that is, Heywood condition [38]; hence, 13 fatty acids remained. Afterwards three dimensions were identified for men and women with eigenvalues greater than one.

3.2. Confirmatory Factor Analysis. Model M0 allowed intercepts, loadings, and unique variances to be freely estimated for men and women. The absolute fit was good with a SRMSR of 0.035, and the fit was much better than a model with zero correlations since CFI = 0.888 (Table 2). However, the fit measures which adjust for parsimony, that is, RMSEA and TLI, were not near acceptable ranges. In model M1 when the 13 fatty acid means were held constant between gender $H_1 : \boldsymbol{\nu}_{male}=\boldsymbol{\nu}_{female}$, the SBC increased by over 400 and the hypothesis was rejected using the chi-squared difference testing between nested models $X^2_{13} = 520$, $P < 0.0001$. As pointed out earlier since robust ML method was used due to

Table 2: Goodness-of-fit for testing gender invariance (among 13 fatty acids).

Model constraint (N = 3196)	Absolute fit					Relative fit	
	χ^2/DF†	χ^2 Scaling	SRMSR	SBC	RMSEA upper 90% limit	CFI	TLI
M0: unrestricted by gender	1900/84	1.079	0.035	100055	0.121	0.888	0.791
M1: equal fatty acid means, $\boldsymbol{\nu}$	2409/97	1.071	0.056	100480	0.126	0.857	0.770
M2: equal loadings, $\boldsymbol{\Lambda}$	1965/114	1.097	0.040	99919	0.105	0.885	0.843
M3: equal unique Variances, $\boldsymbol{\Theta}$	1673/97	1.255	0.036	99999	0.105	0.902	0.843
M4: equal loadings and latent covariances, $\boldsymbol{\Lambda}$,$\boldsymbol{\Psi}$	1986/120	1.105	0.048	99909	0.102	0.884	0.850
M5: equal fatty acid covariance matrix, $\boldsymbol{\Sigma}$	1793/133	1.251	0.050	99853	0.092	0.897	0.879
M6: SEM model	968.8/107	1.217	0.046	98999	0.075	0.947	0.922
M7: reduced SEM M6	967.5/111	1.225	0.046	98972	0.074	0.947	0.925

†The total degrees of freedom (DF) = 2 ∗ (91 fatty acid variances/covariances + 13 fatty acid means) = 208 parameters in all models; hence, the number of estimated parameters equals 208 − X^2 DF; SRMSR = standardized root mean square residual; SBC = Schwarz Bayesian criteria; RMSEA = root mean square error of approximation; CFI = Bentler Comparative Fit Index; TLI = Tucker-Lewis Index.

lack of multivariate normality, the chi-squared test statistic was modified [32]. Specifically, to compare model M1 nested in M0, the scaling correction factor was computed as $c_d = (\mathrm{DF}_{\mathrm{M1}} * \mathrm{Scaling}_{\mathrm{M1}} - \mathrm{DF}_{\mathrm{M0}} * \mathrm{Scaling}_{\mathrm{M0}})/(\mathrm{DF}_{\mathrm{M1}} - \mathrm{DF}_{\mathrm{M0}}) = (97 * 1.071 - 84 * 1.079)/13 = 1.0$. Then $X^2_{13} = (X^2_{\mathrm{M1}} * \mathrm{Scaling}_{\mathrm{M1}} - X^2_{\mathrm{M0}} * \mathrm{Scaling}_{\mathrm{M0}})/c_d = (2409 * 1.071 - 1900 * 1.079)/1.019 = 520$ (Table 2). All other model fit measures deteriorated as well. These results importantly show that the multivariate fatty acid mean profile was not the same between genders for these fatty acids.

Next the fatty acid correlation structure was tested for gender invariance in multiple steps. When comparing model M2 to M0, $\mathrm{H}_2 : \boldsymbol{\Lambda}_{\mathrm{male}} = \boldsymbol{\Lambda}_{\mathrm{female}}$ we concluded that the factor loadings were different between men and women $X^2_{30} = 92$, $P < 0.0001$ (Table 2). To test for gender differences in the fatty acids' variances, $\mathrm{H}_3 : \boldsymbol{\Theta}_{\mathrm{male}} = \boldsymbol{\Theta}_{\mathrm{female}}$ models M3 and M0 were compared. The fatty acids' variances were *not* different between genders, $X^2_{13} = 20.7$, respectively, $P = 0.079$. Next the latent variable covariance structure was tested $\mathrm{H}_4 : \boldsymbol{\Psi}_{\mathrm{male}} = \boldsymbol{\Psi}_{\mathrm{female}}$, by comparing model M2 with M4 and found to be different between men and women ($P < 0.0001$). Likewise the overall covariance structure $\mathrm{H}_5 : \boldsymbol{\Sigma}_{\mathrm{male}} = \boldsymbol{\Sigma}_{\mathrm{female}}$ was different ($P < 0.0001$).

3.3. Structural Equation Modeling Constraints. The above results suggested differential fatty acid functioning for men and women, so separate models were developed by gender that allows the standardized latent variables scores to be compared between men and women. Model M3 had a good fit compared to M0 shown by chi-squared difference testing, SRMSR, and CFI; however, the parsimony measures suggest there were still too many parameters. Model M3 is shown for men and women in Supplemental Tables 3 and 4, respectively; the latent variables were named for the fatty acids with the strongest correlations as PUFA, SATURATED, and TRANS FACTORS. When examining the loading estimates between men and women they were quite similar; there were only 3 parameters that differ by >0.10 which included the saturated fatty acids C14:0 and C18:0. These two fatty acids have slightly stronger correlations with the underlying latent variables in women than men. To further reduce the model complexity, constraints were placed on the loading matrix. A threshold of 0.15 was chosen, and parameters were constrained to zero if they had loadings below this threshold.

Correlations among dietary intakes of fatty acids were determined for the subset of 2332 participants with valid food frequency questionnaires [39] (Supplemental Table 5). Dietary intake (g/d) was available for 11 out of 13 RBC fatty acids included in the latent variable model, and C22:4n6 and C22:5n6 were not calculated from the diet. Since RBC fatty acids were correlated with corresponding dietary intakes of fatty acids, covariances were added to the fatty acids residual matrix, $\boldsymbol{\Theta}$, to account for foods being composed of many different fatty acids. Being able to specify which residual covariances to include is a feature unique to structural equation modeling and cannot be accomplished in the context of confirmatory factor analysis. There were 14 strong dietary correlations ($r > 0.80$) that were added to the model. The correlation between C18:1t and C18:2t was extremely high ($r = 0.98$) and caused model convergence issues; therefore, it was subsequently removed.

The biosynthesis process is well known for omega-3 and omega-6 fatty acids [7]. Delta-6 and delta-5 desaturase activity is needed to convert C18:3n3 into C20:5n3 (Figure 1). Then delta-6 desaturase (D6D) is required again to further convert C20:5n3 into C22:6n3. The amount of D6D available for the second conversion to synthesize C22:6n3 may be

Table 3: Structural equation model M7 factor loadings (Λ) and standardized fatty acid means (ν).

Fatty acids	Men				Women			
	PUFA*	SAT	TRANS	Mean	PUFA*	SAT	TRANS	Mean
Ln(C18:3n3)	0.257	0.138	0.070	−0.186	0.370	0.149	0.114	0.149
Ln(C20:5n3)	0.847	0	0	−0.043	0.844	0	0	0.036
C22:6n3	0.815	−0.373	0	0	0.801	−0.335	0	0
C20:4n6	−0.634	−0.289	−0.215	0	−0.667	−0.285	−0.204	0
C22:4n6	−0.837	0	0	0.168	−0.814	0	0	−0.138
C22:5n6	−0.806	0	0	0.083	−0.788	0	0	−0.069
C14:0	0	0.633	0.056	−0.221	0	0.781	0.105	0.178
C16:0	0	0.754	−0.227	0	0	0.808	−0.290	0
C18:0	0	−0.524	0	0.072	0	−0.681	0.084	−0.054
C16:1	0	0.723	0	−0.209	0	0.812	0	0.169
C16:1t	0	0	0.499	−0.070	0	0	0.540	0.057
C18:1t	0	−0.114	0.888	0	0	−0.168	0.843	0
C18:2t	0	0.248	0.728	−0.065	0	0.239	0.738	0.053
				Factor correlations (Ψ)				
Factor								
PUFA	1		0.190	−0.323	1		0.249	−0.385
SAT	0.190		1	−0.160	0.249		1	0.003
TRANS	−0.323		−0.160	1	−0.385		0.003	1

*Direction of signs is arbitrary; the signs for the PUFA loadings and factor correlations have been reversed in this table to make more n3 fatty acid positively associated with the PUFA FACTOR.

limited by a function of what is initially consumed to support converting C18:3n3 [40]. Therefore, the amount and variability of both C20:5n3 and C22:6n3 depend (to some extent) on the intake of the parent n3 fatty acid C18:3n3. In the omega-6 fatty acid family, D6D is needed for C22:4n6 to synthesize into C22:5n6. These biochemical steps introduce structural elements into the SEM model, so these three covariances were added to fatty acid residuals matrix Θ. However, omega-3 and omega-6 fatty acids also compete for the desaturase enzymes, and the omega-3 fatty acids are the preferential substrates [6]. So with higher levels of C20:5n3 (whether by biosynthesis or fish oil consumption), D5D activity is inhibited (feedback inhibition, whereby the enzyme senses when enough product has been made and then shuts down). This slows the synthesis of C20:4n6 from C20:3n6. Likewise C22:6n3 and C22:5n6 compete for D6D. These two additional fatty acid covariances were added to the SEM residual matrix as well.

3.4. Final Structural Equation Model. After the above constraints were placed on the model, the resulting SEM (Model M6) had a significantly *better* fit than the unrestricted model by gender (M0), $X^2_{23} = 506$ ($P < 0.0001$). Additionally model M6 was the only model to have a "reasonable" fit with RMSEA <0.08. Also it was the only model to have CFI and TLI > 0.90. Model M6 had a total of 208 − 107 = 101 estimated parameters, including the following gender specific 2 ∗ (23 loadings, 9 means, and 3 latent variable covariances) = 70 and gender invariant (13 residual variances and 13 dietary-related and 5 desaturase-related residual covariances) = 31. The loadings and latent variable correlations (Supplemental Table 6) and fatty acid residual matrix (Supplemental Table 7) are shown for model M6. One loading and three residual covariances were <0.05; these were set to zero for a more parsimonious model (M7). The nested fit between the reduced SEM model M7 was similar to model M6, $X^2_4 = 4.3$ ($P = 0.37$) and all the parsimony fit measures (i.e, SBC, RMSEA, and TLI) were improved.

There were four fatty acids with gender mean differences which were not significantly different than zero (i.e, C22:6n3, C20:4n6, C16:0, and C18:1t). The greatest mean differences (all 0.30 to 0.40 SD) between genders were found in two PUFA [ln(C18:3n3) and C22:4n6], one saturated fatty acid [C14:0] and one monounsaturated fatty acid [C16:1]. The loadings, latent variable covariance structure, mean profile, and fatty acid residual matrix for model M7 are shown in Tables 3 and 4. The Mplus code for model M7 is given in Algorithm 1.

4. Discussion

Although 22 individual fatty acids were measured during the GC process, 9 were removed from the latent variable analysis because they were not related to the core structure as explained above. However, the individual fatty acids that were removed may still have clinical utility as individual predictor variables. For example, C18:2n6 (Linoleic acid) was fairly independent of the PUFA, SATURATED, and TRANS latent variable scores, and all correlations were around 0.10 (Supplemental Table 8). Since C18:2n6 has been reported inversely related with heart disease [4], it could still have clinical utility as in independent predictor variable in combination with these newly defined latent

Table 4: Structural equation model M7 fatty acid residuals matrix (**Θ**).

	Ln C18:3n3	Ln C20:5n3	C22:6n3	C20:4n6	C22:4n6	C22:5n6	C14:0	C16:0	C18:0	C16:1	C16:1t	C18:1t	C18:2t
Ln(C18:3n3)	0.832	−0.120*	−0.214*	0	0	0	0	0	0	0	0	0	0
Ln(C20:5n3)	−0.120*	0.278	0.077**	0.081*	0	0	0	0	0	0	0	0	0
C22:6n3	−0.214*	0.077**	0.372	0	0	0.137*	0	0	0	0	0	0	0
C20:4n6	0	0.081*	0	0.466	0	0	0	0	0	0	0	0	0
C22:4n6	0	0	0	0	0.290	0	0	0	0	0	0	0	0
C22:5n6	0	0	0.137*	0	0	0.358	0	0	0	0	0	0	0
C14:0	0	0	0	0	0	0	0.437	−0.066**	0.096**	0	0.090**	0	0
C16:0	0	0	0	0	0	0	−0.066**	0.290	0	0.034**	−0.115**	0	0
C18:0	0	0	0	0	0	0	0.096**	0	0.617	−0.080**	−0.081**	−0.219**	−0.082**
C16:1	0	0	0	0	0	0	0	0.034**	−0.080**	0.360	−0.099**	0	0
C16:1t	0	0	0	0	0	0	0.090**	−0.115**	−0.081**	−0.099**	0.720	0	0
C18:1t	0	0	0	0	0	0	0	0	−0.219**	0	0	0.220	0
C18:2t	0	0	0	0	0	0	0	0	−0.082**	0	0	0	0.426

**indicates dietary intake-related covariances; *indicates biosynthesis-related covariances; fatty acid residual variances are shown on the main diagonal.

```
TITLE:
Structural Equation Model M7;

DATA:
FILE IS infile;

VARIABLE:
NAMES ARE C140 C160 C180 C161 C161t C181t C182t C204n6 C224n6 C225n6
    LnC183n3 LnC205n3 C226n3 ID Female;
USEVARIABLES C140 C160 C180 C161 C161t C181t C182t C204n6 C224n6 C225n6
    LnC183n3 LnC205n3 C226n3;
GROUPING is Female (1=Female 0=Male);
AUXILIARY=ID;

ANALYSIS:
TYPE=GENERAL;
ESTIMATOR=MLR;

MODEL:
!MODEL Female;
!Latent variable loadings;
fFISH BY LnC183n3* LnC205n3 C226n3 C204n6 C224n6 C225n6;
fSAT BY C226n3* C204n6 C140 C160 C180 C161 LnC183n3 C181t C182t;
fTRANS BY C204n6* C160 C180 C161t C181t C182t LnC183n3 C140;

!Latent variable means are fixed at 0;
[fFISH@0 fSAT@0 fTRANS@0];

!Fatty acid means are free, or constrained to zero where indicated;
[LnC183n3 LnC205n3 C226n3@0 C224n6 C225n6 C204n6@0 C140 C180 C160@0 C161 C161t C181t@0 C182t];

!Latent variables covariance matrix;
fFISH@1; fSAT@1; fTRANS@1;
fFISH WITH fSAT;
fFISH WITH fTRANS;
fSAT WITH fTRANS;

!Fatty acid residual variances are equal between gender;
C140(1); C160(2); C180(3); C161(4); C161t(5); C181t(6); C182t(7);
C204n6(8); C224n6(9); C225n6(10); LnC205n3(11); C226n3(12); LnC183n3(13);

!Dietary intake covariances are equal between gender;
C140 WITH C160(14); C140 WITH C180(15); C140 WITH C161t(16);
C160 WITH C161(17); C160 WITH C161t(18);
C180 WITH C161(19); C180 WITH C161t(20); C180 WITH C181t(21); C180 WITH C182t(22);
C161 WITH C161t(23);
LnC205n3 WITH C226n3(24);

!Desaturase enzymes covariances are equal between genders;
LnC183n3 WITH LnC205n3(25);
C204n6 WITH LnC205n3(26);
C226n3 WITH C225n6(27);
C226n3 WITH LnC183n3(28);

MODEL male:
!Latent variable loadings;
fFISH BY LnC183n3* LnC205n3 C226n3 C204n6 C224n6 C225n6;
fSAT BY C226n3* C204n6 C140 C160 C180 C161 LnC183n3 C181t C182t;
fTRANS BY C204n6* C160 C180@0 C161t C181t C182t LnC183n3 C140;

!Fatty acid means are free, or constrained to zero where indicated;
[LnC183n3 LnC205n3 C226n3@0 C224n6 C225n6 C204n6@0 C140 C180 C160@0 C161 C161t C181t@0 C182t];

!Fatty acid residual variances are equal between gender;
C140(1); C160(2); C180(3); C161(4); C161t(5); C181t(6); C182t(7);
C204n6(8); C224n6(9); C225n6(10); LnC205n3(11); C226n3(12); LnC183n3(13);
```

Algorithm 1: Continued.

```
!Dietary intake covariances are equal between gender;
C140 WITH C160(14); C140 WITH C180(15); C140 WITH C161t(16);
C160 WITH C161(17); C160 WITH C161t(18);
C180 WITH C161(19); C180 WITH C161t(20); C180 WITH C181t(21); C180 WITH C182t(22);
C161 WITH C161t(23);
LnC205n3 WITH C226n3(24);

!Desaturase enzymes covariances are equal between genders;
LnC183n3 WITH LnC205n3(25);
C204n6 WITH LnC205n3(26);
C226n3 WITH C225n6(27);
C226n3 WITH LnC183n3(28);

OUTPUT:
SAVEDATA:
FORMAT IS F20.14;
FILE IS outfile;
RESULTS ARE parametersfile;
SAVE=FSCORES;
```

Algorithm 1: Mplus code for structural equation model M7.

variables. C22:5n3 (DPA) is intermediate of C20:5n3 (EPA) and C22:6n3 (DHA) in the biosynthesis process (Figure 1) and was the only excluded fatty acid that had a correlation >0.10 with the PUFA FACTOR ($r = 0.39$). However, DPA has less biological activity than the other marine fish oils [41], so it is not anticipated that DPA would be useful as an individual predictor variable of clinical outcomes. Afterwards, the remaining 13 fatty acids were found to be represented by three dimensions, which constituted nearly 70% of the total fatty acid abundance. In the SEM model, the residual variances were equal between genders. Likewise the structural dietary correlations and biosynthesis processes, which were accounted for with residual correlations, did not vary between men and women. The final SEM model fit the data well by all measures.

The PUFA FACTOR includes the following fatty acids: ln(C18:3n3), ln(C20:5n3), C22:6n3, C20:4n6, C22:4n6, and C22:5n6 (Figure 2); all of which are found in the PUFA biosynthesis processes shown in Figure 1. The loading directions of the fatty acids included in the PUFA FACTOR are also supported by many of the competing metrics being used in fatty acid research. The omega-3 index is implemented in clinical laboratory testing and is defined as RBC C20:5n3 + C22:6n3 [42]. The omega-3 index was an independent risk factor for all-cause mortality in a study of stable coronary heart disease patients, with higher levels indicating reduced risk [3]. Lower amounts of the omega-3 index were associated with depression in a case-control study of adolescents [43]. In the PUFA FACTOR, for both men and women, these two fatty acids operate in the same direction with similar magnitudes, which supports their summation as a biomarker (although C20:5n3 has been log transformed in the PUFA FACTOR).

The n6/n3 ratio [44], n6 HUFA/total HUFA ratio [45], and C20:4n6/C20:5n3 ratio [46] are all metrics that seek to combine individual fatty acids into more powerful predictors of risk. Although the goal is reasonable, these approaches are criticized as being imprecise and impractical [46]. All of these ratios may be flawed in that the same ratio can be obtained by increasing the numerator or decreasing the denominator, when these fatty acids do not have the same physiological properties. An improvement to these ratios may be the PUFA FACTOR. It is a more nuanced metric since it does not simply add up the masses of different PUFA families and create a ratio; it takes into account the relative strengths of relationship among these linearly "opposing" and interrelated fatty acids and reduces this nexus into a single number. The algorithm for scoring these latent variables from raw fatty acid data is given in the following.

Algorithm (algorithm for scoring latent variables)

Step 1. Measure fatty acids as a % of total fatty acids.

Step 2. For C18:3n3 and C20:5n3 transform the values using natural logarithm.

Step 3. Standardize all fatty acids using corresponding overall means and standard deviations from Table 1 into a row vector $\mathbf{Z}_i$ (as shown below, the headings indicate the required order).

Step 4. Calculate the latent variable scores $\boldsymbol{\eta}_i$ for subject i using the appropriate male or female matrices as

$$\boldsymbol{\eta}_i = \left(\mathbf{Z}_i - \boldsymbol{\nu}^T\right)\left[\boldsymbol{\Psi}\boldsymbol{\Lambda}^T\left(\boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}^T + \boldsymbol{\Theta}\right)^{-1}\right]^T, \tag{1}$$

where $\boldsymbol{\nu}$ is the standardized fatty acid mean column vector given by gender in Table 3. $\boldsymbol{\Psi}$ is the latent variable covariance matrix given by gender in Table 3. $\boldsymbol{\Lambda}$ is the loading matrix given by gender in Table 3. $\boldsymbol{\Theta}$ is the fatty acid residual

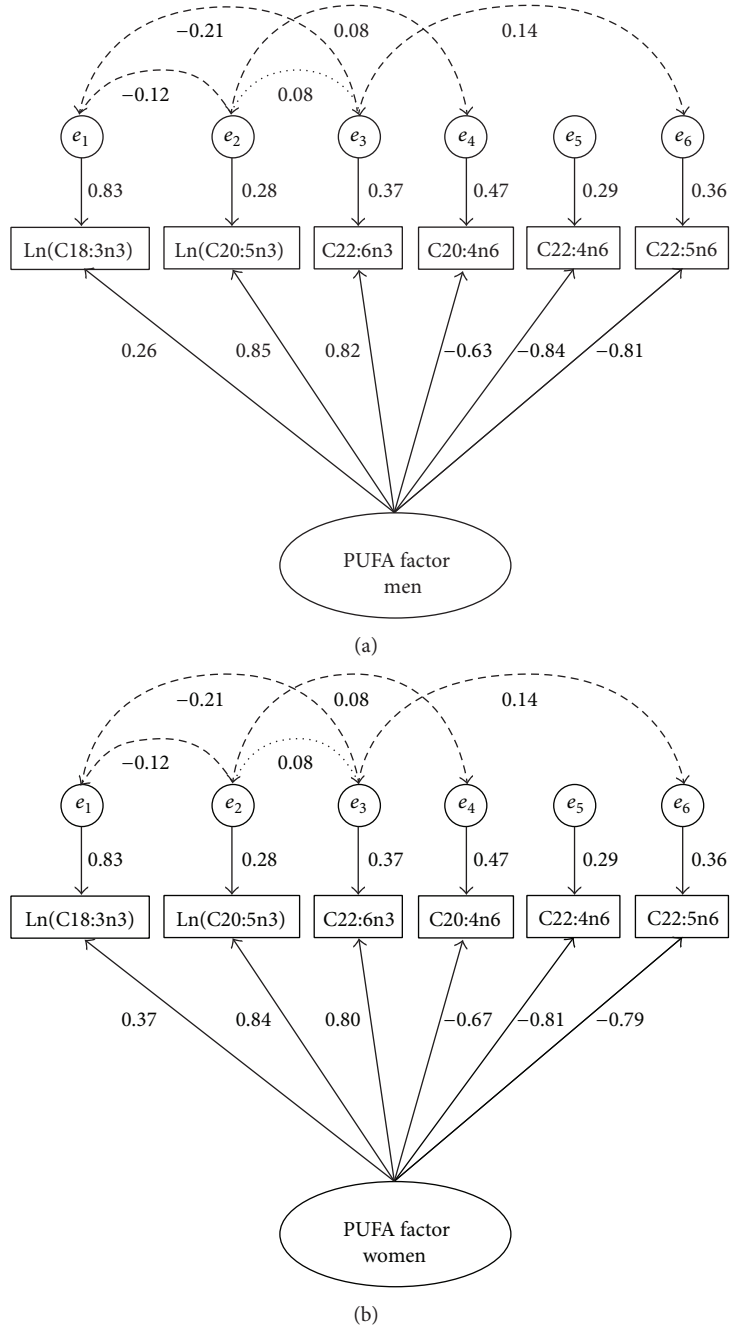


Figure 2: Structural equation model M7 for men (a) and women (b). Solid lines from PUFA FACTOR to fatty acids are gender specific *loadings*, solid lines from circles to fatty acids are residual variances, dotted line indicates structural dietary intake correlation, and dashed lines indicate structural desaturase enzymes required for biosynthesis.

covariance matrix given in Table 4. T means to transpose the matrix. −1 means to take the inverse of the matrix.

Example. Measure RBC fatty acid percent weight composition using gas chromatography as detailed in Harris et al. [41] or similar, then log transform C18:3n3 and C20:5n3, and then standardize all raw data by (value − mean)/SD from Table 1 to produce a row vector:

$$\mathbf{Z}_i = \begin{matrix} \text{Ln} \\ \text{C18:3n3} \end{matrix} \begin{matrix} \text{Ln} \\ \text{C20:5n3} \end{matrix} \begin{matrix} \\ \text{C22:6n3} \end{matrix} \begin{matrix} \\ \text{C20:4n6} \end{matrix} \begin{matrix} \\ \text{C22:4n6} \end{matrix} \begin{matrix} \\ \text{C22:5n6} \end{matrix} \begin{matrix} \\ \text{C14:0} \end{matrix} \begin{matrix} \\ \text{C16:0} \end{matrix} \begin{matrix} \\ \text{C18:0} \end{matrix} \begin{matrix} \\ \text{C16:1} \end{matrix} \begin{matrix} \\ \text{C16:1t} \end{matrix} \begin{matrix} \\ \text{C18:1t} \end{matrix} \begin{matrix} \\ \text{C18:2t} \end{matrix}$$
$$|0.557 \quad 1.776 \quad 2.532 \quad -2.086 \quad -2.495 \quad -1.969 \quad 0.417 \quad 2.248 \quad -2.526 \quad 0.388 \quad -0.213 \quad -1.267 \quad 0.398|. \tag{2}$$

Lastly the latent variable scores are derived by using $\boldsymbol{\eta}_i = (\mathbf{Z}_i - \boldsymbol{\nu}^T)[\boldsymbol{\Psi}\boldsymbol{\Lambda}^T(\boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}^T + \boldsymbol{\Theta})^{-1}]^T$, with the appropriate vector and matrices used for men or women given in Tables 3 and 4. If the blood sample was from a man or woman, the respective latent variables scores would be

$$\begin{array}{rccc} & \text{PUFA FACTOR} & \text{SAT FACTOR} & \text{TRANS FACTOR} \\ \boldsymbol{\eta}_i = & |2.77 & 1.76 & -1.56| \end{array}. \quad (3)$$

$$\begin{array}{rccc} \text{or} & \text{PUFA FACTOR} & \text{SAT FACTOR} & \text{TRANS FACTOR} \\ \boldsymbol{\eta}_i = & |2.56 & 1.38 & -1.52| \end{array}.$$

Another approach has been to construct "desaturase ratios" which are based on the known biosynthetic relationships among PUFA [7]. Since it is far too invasive (requiring liver biopsy) to measure the activity of these enzymes directly, they have been estimated empirically by dividing RBC levels of product fatty acids by levels of precursor fatty acids. Thus the delta-6 desaturase (D6D) activity can be estimated by the ratio of 20:3n6/C18:2n6 and the delta-5 desaturase (D5D) activity by the ratio of C20:4n6/C20:3n6 (Figure 1). Interestingly, both of these desaturase ratios have been associated with risk for the development of Type 2 diabetes mellitus in a recent metareview [47]. The ultimate clinical utility of the PUFA FACTOR (versus desaturase or other fatty acid ratios) will be determined in future studies by comparing these metrics as predictors of disease outcomes for mortality, CHD events, development of type 2 diabetes or dementia, and so forth.

5. Conclusion

The PUFA FACTOR has much supporting evidence based on fatty acid metabolism and dietary patterns. It was also the first dimension extracted from the data, due to explaining the most variability (about 30% of total in men and women) for these 13 fatty acids. In a previous study these same Framingham subjects were included in a heritability analysis, and it was found that about 25% and 40% of the variance in two of the fatty acids included in the PUFA factor (i.e, EPA and DHA) was due to genetic and environment, respectively [41]. The PUFA factor can also be seen as a unifying theme among the various n3 and n6 metrics typically used in fatty acid research. Since n3 and n6 fatty acids have been implicated in cardiovascular diseases [4, 47], cognitive function [48], brain magnetic resonance imaging [49, 50], depression [43], mortality [1–3], and cellular aging [51] it is reasonable to expect the PUFA FACTOR to have clinical utility for predicting these outcomes. In contrast, the SATURATED and TRANS FACTORS had several cross loadings between them and even include some PUFA. Thus, their interpretations are unclear, which will likely limit their usefulness.

The strengths of this study include a well-characterized structural equation model applied to RBC fatty acid data which incorporates elements of both fatty acid metabolism and dietary intake patterns in defining the model. Additionally the correlation structure of the SEM was decomposed, and the separate components were tested for gender invariance. Another benefit was the use of a large, extensively studied cohort with enrichment for minorities (Framingham). Limitations include that the RBC measurements were from a particular GC method, and since national standards have not been established for measuring fatty acids the sensitivity of these results to other GC methods is unknown. This study measured erythrocytes; other blood fractions or sample types (e.g., whole blood, plasma, and plasma phospholipids) have different rank orders of fatty acid abundances and these may require unique structural equation models. The fit of this SEM needs be tested in independent samples to determine its generalizability beyond the Framingham Study.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors would like to acknowledge Ramachandran S. Vasan and Sander J. Robins at the Framingham Heart Study for their assistance in developing the overall research plan. The authors are supported in part by the National Heart Lung and Blood Institute (NHLBI; R01HL089590) and by contract N01-HC-25195, the Framingham Heart Study (NHLBI) and Boston University School of Medicine.

References

[1] C. M. Albert, H. Campos, M. J. Stampfer et al., "Blood levels of long-chain n-3 fatty acids and the risk of sudden death," *The New England Journal of Medicine*, vol. 346, no. 15, pp. 1113–1118, 2002.

[2] C. Wang, W. S. Harris, M. Chung et al., "n-3 Fatty acids from fish or fish-oil supplements, but not α-linolenic acid, benefit cardiovascular disease outcomes in primary- and secondary-prevention studies: a systematic review," *American Journal of Clinical Nutrition*, vol. 84, no. 1, pp. 5–17, 2006.

[3] J. V. Pottala, S. Garg, B. E. Cohen, M. A. Whooley, and W. S. Harris, "Blood eicosapentaenoic and docosahexaenoic acids predict all-cause mortality in patients with stable coronary heart disease: the heart and soul study," *Circulation: Cardiovascular Quality and Outcomes*, vol. 3, no. 4, pp. 406–412, 2010.

[4] W. S. Harris, D. Mozaffarian, E. Rimm et al., "Omega-6 fatty acids and risk for cardiovascular disease: a science advisory from the American Heart Association nutrition subcommittee of the council on nutrition, physical activity, and metabolism; council on cardiovascular nursing; and council on epidemiology and prevention," *Circulation*, vol. 119, no. 6, pp. 902–907, 2009.

[5] N. T. Bendsen, S. Stender, P. B. Szecsi et al., "Effect of industrially produced trans fat on markers of systemic inflammation: evidence from a randomized trial in women," *Journal of Lipid Research*, vol. 52, no. 10, pp. 1821–1828, 2011.

[6] J. B. Barham, M. B. Edens, A. N. Fonteh, M. M. Johnson, L. Easter, and F. H. Chilton, "Addition of eicosapentaenoic acid to γ-linolenic acid-supplemented diets prevents serum

arachidonic acid accumulation in humans," *Journal of Nutrition*, vol. 130, no. 8, pp. 1925–1931, 2000.

[7] J. R. Marszalek and H. F. Lodish, "Docosahexaenoic acid, fatty acid-interacting proteins, and neuronal function: breastmilk and fish are good for you," *Annual Review of Cell and Developmental Biology*, vol. 21, pp. 633–657, 2005.

[8] D. Mozaffarian, A. Ascherio, F. B. Hu et al., "Interplay between different polyunsaturated fatty acids and risk of coronary heart disease in men," *Circulation*, vol. 111, no. 2, pp. 157–164, 2005.

[9] R. J. T. Mocking, J. Assies, A. Lok et al., "Statistical methodological issues in handling of fatty acid data: percentage or concentration, imputation and indices," *Lipids*, vol. 47, no. 5, pp. 541–547, 2012.

[10] C. K. Chow, "Fatty acid composition of plasma phospholipids and risk of prostate cancer," *The American Journal of Clinical Nutrition*, vol. 89, no. 6, pp. 1946–1947, 2009.

[11] F. L. Crowe, "Reply to CK Chow," *American Journal of Clinical Nutrition*, vol. 89, no. 6, pp. 1946–1947, 2009.

[12] K. E. Bradbury, C. Murray Skeaff, T. J. Green, A. R. Gray, and F. L. Crowe, "The serum fatty acids myristic acid and linoleic acid are better predictors of serum cholesterol concentrations when measured as molecular percentages rather than as absolute concentrations," *American Journal of Clinical Nutrition*, vol. 91, no. 2, pp. 398–405, 2010.

[13] H. Harris, "Omega-3 fatty acids in cardiac biopsies from heart transplantation patients—correlation with erythrocytes and response to supplementation," *Circulation*, vol. 110, no. 12, pp. 1645–1649, 2004.

[14] L. Hodson, C. M. Skeaff, and B. A. Fielding, "Fatty acid composition of adipose tissue and blood in humans and its use as a biomarker of dietary intake," *Progress in Lipid Research*, vol. 47, no. 5, pp. 348–380, 2008.

[15] M. A. Beydoun, J. S. Kaufman, J. Ibrahim, J. A. Satia, and G. Heiss, "Measurement error adjustment in essential fatty acid intake from a food frequency questionnaire: alternative approaches and methods," *BMC Medical Research Methodology*, vol. 7, article 41, 2007.

[16] M. Tournoud, R. Ecochard, J. Iwaz, J.-P. Steghens, G. Bellon, and I. Durieu, "Structural equations to model relationships between pulmonary function, fatty acids and oxidation in cystic fibrosis," *Scandinavian Journal of Clinical and Laboratory Investigation*, vol. 69, no. 1, pp. 36–44, 2009.

[17] J. C. N. Chan, P. C. Y. Tong, and J. A. J. H. Critchley, "The insulin resistance syndrome: mechanisms of clustering of cardiovascular risk," *Seminars in Vascular Medicine*, vol. 2, no. 1, pp. 45–57, 2002.

[18] J. Bradbury, L. Brooks, and S. P. Myers, "Are the adaptogenic effects of omega 3 fatty acids mediated via inhibition of proinflammatory cytokines?" *Evidence-Based Complementary and Alternative Medicine*, vol. 2012, Article ID 209197, 14 pages, 2012.

[19] P. W. F. Wilson, R. B. D'Agostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel, "Prediction of coronary heart disease using risk factor categories," *Circulation*, vol. 97, no. 18, pp. 1837–1847, 1998.

[20] H. W. Marsh, B. Muthén, T. Asparouhov et al., "Exploratory structural equation modeling, integrating CFA and EFA: application to students' evaluations of university teaching," *Structural Equation Modeling*, vol. 16, no. 3, pp. 439–476, 2009.

[21] H. W. Marsh, O. Lüdtke, B. Muthén et al., "A new look at the big five factor structure through exploratory structural equation modeling," *Psychological Assessment*, vol. 22, no. 3, pp. 471–491, 2010.

[22] W. B. Kannel, M. Feinleib, and P. M. McNamara, "An investigation of coronary heart disease in families. The Framingham offspring study," *American Journal of Epidemiology*, vol. 110, no. 3, pp. 281–290, 1979.

[23] S. F. Quan, B. V. Howard, C. Iber et al., "The Sleep Heart Health Study: design, rationale, and methods," *Sleep*, vol. 20, no. 12, pp. 1077–1085, 1997.

[24] W. S. Harris, J. V. Pottala, R. S. Vasan et al., "Changes in erythrocyte membrane trans and marine fatty acids between 1999 and 2006 in older Americans," *Journal of Nutrition*, vol. 142, no. 7, pp. 1297–1303, 2012.

[25] A. Skrondal and S. Rabe-Hesketh, *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Modeling*, Chapman & Hall, Boca Raton, Fla, USA, 2004.

[26] R. B. Kline, *Principles and Practice of Structural Equation Modeling*, The Guilford Press, New York, NY, USA, 2nd edition, 2005.

[27] K. V. Mardia, "Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies," *Sankhya: The Indian Journal of Statistics B*, vol. 36, pp. 115–128, 1974.

[28] L. K. Muthen and B. O. Muthen, *Mplus User's Guide*, Muthen & Muthen, Los Angeles, Calif, USA, 6th edition, 2010.

[29] H. F. Kaiser, "An index of factorial simplicity," *Psychometrika*, vol. 39, no. 1, pp. 31–36, 1974.

[30] M. A. Pett, N. R. Lackey, and J. J. Sullivan, *Making Sense of Factor Analysis: The Use of Factor Analysis for Instrument Development in Health Care Research*, Sage Publications, Thousand Oaks, Calif, USA, 2003.

[31] R. Johnson and D. Wichern, *Applied Multivariate Statistical Analysis*, Pearson Education, Upper Saddle River, NJ, USA, 6th edition, 2007.

[32] A. Satorra and P. M. Bentler, "A scaled difference chi-square test statistic for moment structure analysis," *Psychometrika*, vol. 66, no. 4, pp. 507–514, 2001.

[33] J. H. Steiger and J. M. Lind, "Statistically based tests for the number of common factors," in *Proceedings of the Annual Meeting Psychometric Society*, Iowa City, IA, Iowa, USA, 1980.

[34] P. M. Bentler, "Comparative fit indexes in structural models," *Psychological Bulletin*, vol. 107, no. 2, pp. 238–246, 1990.

[35] G. W. Cheung and R. B. Rensvold, "The effects of model parsimony and sampling error on the fit of structural equation models," *Organizational Research Methods*, vol. 4, no. 3, pp. 236–264, 2001.

[36] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[37] H. Akaike, "Information theory and an extension of maximum likelihood principal," in *Proceedings of the 2nd International Symposium of Information Theory and Control*, Akademia Kiado, Budapest, Hungary, 1973.

[38] F. Chen, K. A. Bollen, P. Paxton, P. J. Curran, and J. B. Kirby, "Improper solutions in structural equation models: causes, consequences, and strategies," *Sociological Methods and Research*, vol. 29, no. 4, pp. 468–508, 2001.

[39] M. E. Rumawas, J. T. Dwyer, N. M. Mckeown, J. B. Meigs, G. Rogers, and P. F. Jacques, "The development of the mediterranean-style dietary pattern score and its application to the american diet in the framingham offspring cohort," *Journal of Nutrition*, vol. 139, no. 6, pp. 1150–1156, 2009.

[40] R. Portolesi, B. C. Powell, and R. A. Gibson, "Competition between 24:5n-3 and ALA for Δ6 desaturase may limit the accumulation of DHA in HepG2 cell membranes," *Journal of Lipid Research*, vol. 48, no. 7, pp. 1592–1598, 2007.

[41] W. S. Harris, J. V. Pottala, S. M. Lacey et al., "Clinical correlates and heritability of erythrocyte eicosapentaenoic and docosahexaenoic acid content in the Framingham Heart Study," *Atherosclerosis*, vol. 225, no. 2, pp. 425–431, 2012.

[42] W. S. Harris and C. Von Schacky, "The Omega-3 Index: a new risk factor for death from coronary heart disease?" *Preventive Medicine*, vol. 39, no. 1, pp. 212–220, 2004.

[43] J. V. Pottala, J. A. Talley, S. W. Churchill, D. A. Lynch, C. von Schacky, and W. S. Harris, "Red blood cell fatty acids are associated with depression in a case-control study of adolescents," *Prostaglandins Leukotrienes and Essential Fatty Acids*, vol. 86, no. 4-5, pp. 161–165, 2012.

[44] A. P. Simopoulos, "The importance of the omega-6/omega-3 fatty acid ratio in cardiovascular disease and other chronic diseases," *Experimental Biology and Medicine*, vol. 233, no. 6, pp. 674–688, 2008.

[45] W. E. M. Lands, "Diets could prevent many diseases," *Lipids*, vol. 38, no. 4, pp. 317–321, 2003.

[46] W. S. Harris, "The omega-6/omega-3 ratio and cardiovascular disease risk: uses and abuses," *Current Atherosclerosis Reports*, vol. 8, no. 6, pp. 453–459, 2006.

[47] J. Kröger and M. B. Schulze, "Recent insights into the relation of Δ5 desaturase and Δ6 desaturase activity to the development of type 2 diabetes," *Current Opinion in Lipidology*, vol. 23, no. 1, pp. 4–10, 2012.

[48] J. G. Robinson, N. Ijioma, and W. Harris, "Omega-3 fatty acids and cognitive function in women," *Women's Health (London, England)*, vol. 6, no. 1, pp. 119–134, 2010.

[49] J. V. Pottala, K. Yaffe, J. G. Robinson et al., "Higher RBC EPA + DHA corresponds with larger total brain and hippocampal volumes: WHIMS-MRI Study," *Neurology*, vol. 82, no. 5, pp. 435–442, 2014.

[50] Z. S. Tan, W. S. Harris, A. S. Beiser et al., "Red blood cell omega-3 fatty acid levels and markers of accelerated brain aging," *Neurology*, vol. 78, no. 9, pp. 658–664, 2012.

[51] R. Farzaneh-Far, J. Lin, E. S. Epel, W. S. Harris, E. H. Blackburn, and M. A. Whooley, "Association of marine omega-3 fatty acid levels with telomeric aging in patients with coronary heart disease," *JAMA—Journal of the American Medical Association*, vol. 303, no. 3, pp. 250–257, 2010.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 242717, 7 pages
http://dx.doi.org/10.1155/2014/242717

Research Article

Use of CHAID Decision Trees to Formulate Pathways for the Early Detection of Metabolic Syndrome in Young Adults

Brian Miller,[1] Mark Fridline,[2] Pei-Yang Liu,[1] and Deborah Marino[1]

[1] *School of Nutrition and Dietetics, College of Health Professions, The University of Akron, Akron, OH 44325-6102, USA*
[2] *Department of Statistics, College of Arts and Sciences, University of Akron, Akron, OH 44325-1913, USA*

Correspondence should be addressed to Mark Fridline; mmf@uakron.edu

Received 21 January 2014; Accepted 16 March 2014; Published 10 April 2014

Academic Editor: Zhenyu Jia

Metabolic syndrome (MetS) in young adults (age 20–39) is often undiagnosed. A simple screening tool using a surrogate measure might be invaluable in the early detection of MetS. *Methods*. A chi-squared automatic interaction detection (CHAID) decision tree analysis with waist circumference user-specified as the first level was used to detect MetS in young adults using data from the National Health and Nutrition Examination Survey (NHANES) 2009-2010 Cohort as a representative sample of the United States population ($n = 745$). *Results*. Twenty percent of the sample met the National Cholesterol Education Program Adult Treatment Panel III (NCEP) classification criteria for MetS. The user-specified CHAID model was compared to both CHAID model with no user-specified first level and logistic regression based model. This analysis identified waist circumference as a strong predictor in the MetS diagnosis. The accuracy of the final model with waist circumference user-specified as the first level was 92.3% with its ability to detect MetS at 71.8% which outperformed comparison models. *Conclusions*. Preliminary findings suggest that young adults at risk for MetS could be identified for further followup based on their waist circumference. Decision tree methods show promise for the development of a preliminary detection algorithm for MetS.

1. Introduction

Metabolic Syndrome (MetS) is a collection of cardiometabolic risk factors that includes excessive central adiposity, elevated triglycerides (TG) and fasting plasma glucose (FPG), decreased HDL-cholesterol (HDL), and hypertension [1]. When these risk factors are present in tandem, they increase the risk of heart attack, stroke, and cardiovascular morbidity and/or mortality affecting one in three adults in the United States (US) [2]. Additionally, there is a disproportionate increase in healthcare costs for adults presenting with MetS compared to those that do not [3, 4]. Prevalence and complications associated with MetS and other cardiometabolic diseases continue to be a major health concern in the United States.

The National Cholesterol Education Program Adult Treatment Panel III (NCEP) and International Diabetes Federation (IDF) clinical risk models are limited in their usefulness in that they only identify either the presence or absence of MetS [3, 4]. However much like obesity, there are varied clinical implications based on the severity of the risk factors used to define MetS. Furthermore, certain factors might be more significant than others in predicting the presence or absence of MetS. Waist circumference has been demonstrated to be a strong predictor of cardiometabolic risk [2, 5, 6] and can be easily and affordably obtained in a clinical screening. Creating an early detection model that stratifies the severity of cardiometabolic and anthropometric factors used in the MetS diagnosis based on proxy measures easily obtained in a clinical setting would be invaluable for clinicians aiming to provide improved patient-centered care [7].

Predictive models are useful and cost effective in identifying risk of developing cardiometabolic chronic diseases [8]. Decision tree methodologies show promise over traditional predictive modeling procedures based on their ease of interpretability by nonstatisticians. One of the outstanding advantages of decision tree analysis is that it can visualize

the relationship pathways between the binary target variable and the related continuous and/or categorical predictor variables with a tree image [9]. Recently, Worachartcheewan et al. [10] used a classification and regression tree (CART) model to identify pathways for MetS detection in accordance with the NCEP criteria using a large Thai population of overweight men and women without regard to age or health status. In this model, TG was the strongest predictor of MetS. However, dyslipidemia is not commonly elevated in younger adulthood and is invasive and costly to measure [7].

Unfortunately, there is a lack of research focusing on the adult population ages of 20–39 years where preventative or early corrective measures can be utilized. Rather, the majority of research has focused on the adult population greater than age 40 [11, 12]. Currently no preventative methodologies exist for the early detection of MetS. Therefore, attention is warranted to the derivation of premetabolic syndrome criteria that identifies at-risk subjects who can utilize preventive intervention well before qualifying as moderate to high risk on current predictive models [13].

The purpose of this pilot study is to investigate the utility of the chi-squared automatic interaction detection (CHAID) algorithm to identify and develop pathways for the early detection of MetS. The central hypothesis states that the decision tree pathways derived from CHAID algorithms using data from National Health and Nutrition Examination Survey (NHANES) 2009-2010 will detect the presence of MetS in adults of 20–39 years of age. These pathways are meant to serve as pilots for the future development of an easily interpreted, clinically relevant, cost-effective screening tool to detect cardiometabolic chronic disease [14].

2. Materials and Methods

2.1. Participants. The current study is based on publicly available data from the National Health and Nutrition Examination Survey (NHANES) 2009-2010 cohort [15]. The full data set includes 10,537 subjects designed to represent the population of the United States across age, sex, and ethnicity. Subjects with missing MetS criteria were excluded from the present study due to the inability in making a complete classification of MetS (subjects lost $n = 7589$). Subjects not meeting the inclusion criteria of an age between 20 and 39 years were excluded as were those with a body mass index (BMI) less than 20 kg/m^2 (subjects lost $n = 2203$; $n = 522$ for age <20 years, $n = 1622$ for age >39 years, and $n = 59$ for BMI <20 kg/m^2). The final sample retained meeting the inclusion criteria included 745 subjects.

Demographic information included age, sex, and dichotomous ethnicity represented as ethnic or nonethnic. Anthropometric information included weight (kg), height (cm), BMI (kg/m^2), and waist circumference (cm). Laboratory measures included HDL (mg/dl), TG (mg/dl), fasting plasma glucose (FPG, mg/dl), and blood pressure expressed as systolic and diastolic pressures (mmHg).

The criteria for MetS followed the NCEP guidelines defined as presenting with three or more of the following factors: waist circumference > 88 cm for women or >102 cm for men, blood pressure ≥ 135/≥85 mmHg, TG ≥ 150 mg/dl, HDL < 50 mg/dl for women or <40 mg/dl for men, or FPG ≥ 100 mg/dl [16]. Sample characteristics are illustrated in Table 1 and are expressed as mean ± standard deviation. Of the 745 subjects between the ages of 20–39 years, 20% ($n = 149$) presented with the NCEP criteria for MetS. Approval for this analysis was provided by the University of Akron Institutional Review Board.

2.2. Statistical Analysis. The data was arranged in a column-wise format with each subject given a sequence identifier. Data management was performed using data set merging and data subset functions with statistical analysis performed using IBM SPSS version 19. A CHAID algorithm analysis was used to develop the decision tree models. CHAID decision trees are nonparametric procedures that make no assumptions of the underlying data. This algorithm determines how continuous and/or categorical independent variables best combine to predict a binary outcome based on "if-then" logic by portioning each independent variable into mutually exclusive subsets based on homogeneity of the data. For this study, the response variable is the presence or absence of MetS. According to Kass (1980), the CHAID algorithm operates using a series of merging, splitting, and stopping steps based on user-specified criteria as follows [17].

The merging step operates using each predictor variable where CHAID merges nonsignificant categories using the following algorithm.

(1) Perform cross-tabulation of the predictor variable with the binary target variable.

(2) If the predictor variable has only 2 categories, go to step 6.

(3) χ^2-test for independence is performed for each pair of categories of the predictor variable in relation to the binary target variable using the χ^2 distribution (df = 1) with significance (α_{merge}) set at 0.05. For nonsignificant outcomes, those paired categories are merged.

(4) For nonsignificant tests identified by $\alpha_{\text{merge}} > 0.05$, those paired categories are merged into a single category. For tests reaching significance identified by $\alpha_{\text{merge}} \leq 0.05$, the pairs are not merged.

(5) If any category has less than the user-specified minimum segment size, that pair is merged with the most similar other category.

(6) The adjusted P value for the merged categories using a Bonferroni adjustment is utilized to control for Type I error rate.

The splitting step occurs following the determination of all the possible merges for each predictor variable. This step selects which predictor is to be used to "best" split the node using the following algorithm.

(1) χ^2-test for independence using an adjusted P value for each predictor.

Table 1: Subject demographics and descriptive statistics.

Parameter	Mean ± standard deviation ($n = 745$)
Age (yr)	29.3 ± 5.8
Weight (kg)	82.7 ± 21.3
Height (cm)	168.2 ± 9.9
Body mass index (kg/m^2)	29.2 ± 6.8
Systolic blood pressure (mmHg)	113.8 ± 11.7
Diastolic blood pressure (mmHg)	66.7 ± 11.8
Waist circumference (cm)	96.8 ± 15.8
HDL (mg/dl)	51.30 ± 14.9
Triglyceride (mg/dl)	126.7 ± 114.3
Fasting plasma glucose (mg/dl)	98.0 ± 24.6

Values are mean ± standard deviation. HDL: high-density lipoprotein cholesterol ($n = 745$; male = 335, female = 410).

(2) The predictor with the smallest adjusted P value (i.e., most statistically significant) is split if the P value less than the user-specified significance split level (α_{split}) is set at 0.05; otherwise the node is not split and is then considered a terminal node.

The stopping step utilizes the following user-specified stopping rules to check if the tree growing process should stop.

(1) If the current tree reached the maximum tree depth level, the tree process stops.

(2) If the size of a node is less than the user-specified minimum node size, the node will not be split.

(3) If the split of a node results in a child node whose node size is less than the user-specified minimum child node size value, the node will not be split. The parent node is the level where the data set divides into child nodes that can themselves become either parent nodes or end in a terminal or decision node.

(4) The CHAID algorithm will continue until all the stopping rules are met.

The CHAID analysis was run in duplicate with parent nodes defined at 20 subjects, child node defined at 5 subjects, and significance set at (α_{merge}, α_{split}, and P value) ≤ 0.05.

For the first run, the first level or first division was user-specified as waist circumference due to the measurement of this parameter having the lowest cost in MetS screening [18, 19]. The second run was utilized as a comparison to the first model with no first division user-specified. This allowed the algorithm to determine the parameter of the first split. CHAID accuracy and detection was expressed as percentages.

Logistic regression with testing for multicollinearity was performed on the five factors used to define MetS as a parametric comparison to the CHAID models. Results were expressed as overall accuracy of the logistic regression model and detection of MetS, both expressed as percentages with significance of the overall model set at $P \leq 0.05$.

3. Results

3.1. CHAID: Waist Circumference User-Specified. The decision tree algorithm partitioned the data into statistically significant subgroups that were mutually exclusive and exhaustive [17]. The tree analysis in Figure 1 shows the 4-level CHAID tree with a total of 29 nodes, of which 15 were terminal nodes. Four major predictor variables reached significance to be included in this model including waist circumference, TG, HDL, and FPG. The blood pressure MetS criteria, sex, age, and ethnicity did not reach significance for inclusion in the model. This model had an overall classification accuracy of 92.3% with its ability to detect MetS at 71.8%.

The first level of the tree was split into four initial branches according to the user-specified first level on waist circumference. The mean waist circumference of this sample was 96.82 cm with 49.1% of the total population and 86.6% of the population with MetS presenting with the NCEP waist circumference criteria. The MetS prevalence of subjects whose waist circumference was less than 86 cm was 0.5%, which was significantly less than subjects whose waist circumference was between 86 and 94 cm, between 94 and 103 cm, or greater than 103 cm (8.8%, 21.5%, or 45.8%, resp.).

As seen in the second level of the tree, HDL and TG were shown to be the next best predictor variables for each of the waist circumference splits in the first level. The subset of subjects categorized by a waist circumference less than 86 cm and who had HDL less than or equal to 38 mg/dl had a higher prevalence of MetS (4.8%) than those who had an HDL greater than 38 mg/dl (0.0%). In the subset of subjects with a waist circumference greater than 103 cm, the next split based on HDL of less than or equal to 38 mg/dl and 38–49 mg/dl and greater than 49 mg/dl had MetS prevalence of 82.1%, 45.3%, and 7.9%, respectively.

In the subset of subjects categorized by a waist circumference between 86 and 94 cm the next level based on TG less than 138 mg/dl resulted in lower MetS prevalence (1.6%) compared to TG greater than 138 mg/dl (36.4%). The subset of subjects categorized by a waist circumference between 94 and 103 cm and the next level of TG greater than 162 mg/dl had a MetS prevalence of 57.8% compared to TG less than or equal to 162 mg/dl (5.1% MetS). These results indicate that further testing for MetS might not be warranted for subjects presenting with a waist circumference less than 86 cm but would be recommended for those in either of the subcategories of waist circumference.

FPG level was the most prominent variable in the third level of the tree. The only exception was the split based HDL for subjects who had a waist circumference between 94 and 103 cm and TG level was greater than 162 mg/dl. In the subset of subjects whose waist circumference was between 86 and 94 cm and TG level was less than or equal to 138 mg/dl, FPG less than or equal to 103 mg/dl resulted in 0% MetS prevalence compared to the subset greater than 103 mg/dl (16.7%). This was consistent for subjects who had TG greater than 138 mg/dl with the next level based on FPG less than or equal to 92 mg/dl (0%) compared to FPG greater than 92 mg/dl (52.2%).

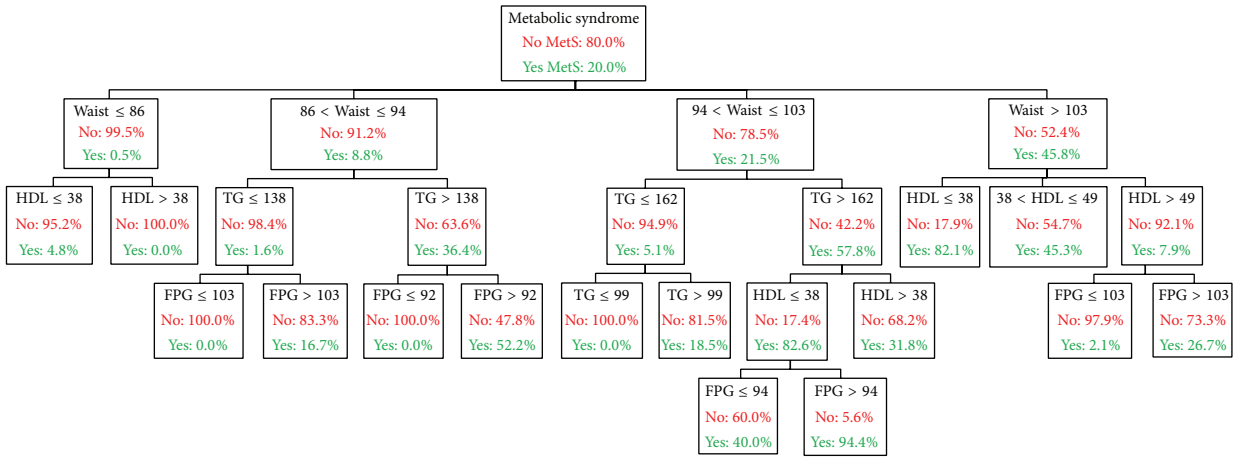


Figure 1: MetS: metabolic syndrome, TG: triglyceride (mg/dl), HDL: high-density lipoprotein cholesterol (mg/dl), Waist: waist circumference (cm), and FPG: fasting plasma glucose (mg/dl).

In the subset of subjects whose waist circumference was between 94 and 103 cm and TG level was less than or equal to 162 mg/dl, FPG again resulted in a 0% MetS prevalence compared to FPG greater than 162 mg/dl (18.5%). In the subset of subjects whose waist circumference was between 94 and 103 cm and TG level was greater than 162 mg/dl, HDL less than or equal to 38 mg/dl resulted in higher MetS prevalence of 82.6% compared to HDL greater than 38 mg/dl (31.8%). In the subset of subjects whose waist circumference was greater than 103 cm, HDL level greater than 49 mg/dl and FPG greater than 103 mg/dl had a MetS prevalence of 2.1% compared to FPG less than or equal to 103 mg/dl (26.7%). Note that FPG level was the only variable in the fourth level of the tree.

Terminal nodes (nodes that do not split any further) are the ends of each pathway where the prevalence is equated to the likelihood of presenting with MetS. Decision rules for the detection of MetS, presented in Table 2, show the "if-then" logic for each of the 15 terminal nodes. The terminal nodes are chronologically sorted by the proportion of Mets detected, where the highest proportion of 94.4% MetS occurred in node 29 and the lowest proportion of 0% occurred in nodes 6, 14, 16, and 18.

3.2. Model Comparison. The following are the results of the user-specified first split model, referred to as the proposed CHAID model, as compared to the CHAID model with no user-specified first split and a logistic regression derived model.

For the CHAID model with no user-specified first split, the first variable was split on FPG. Like the proposed CHAID model, four major predictor variables were selected by the algorithm in this model including waist circumference, TG, HDL, and FPG. The blood pressure MetS criteria, age, sex, and ethnicity did not reach significance and thus were not used in the model. Compared to the proposed CHAID model, this model had a lower, but not practically different, overall classification accuracy of 92.2% with its ability to detect MetS at 69.8%.

The logistic regression model based on the MetS criteria used in CHAID models had no violations of multicollinearity with the model reaching significance. Compared to proposed CHAID model, this logistic regression model had a lower overall classification accuracy of 89.4% with its ability to detect MetS at 61.7%.

4. Discussion

The current study aimed to generate a model for the early detection of MetS in young adults. This model was derived using a CHAID algorithm based on the presence of MetS as the target variable and the MetS classification criteria as its predictors whose values were obtained from 2009-2010 NHANES data. MetS is classified by the presence of 3 of 5 criteria defined by either the NCEP or IDF guidelines. The novelty of this study is that the pathways derived from this model show promise in accurately detecting MetS with an easily obtained measurement.

The CHAID model illustrates multilevel interactions among risk factors to identify stepwise pathways to detect MetS. The five variables (waist circumference, TG, HDL, FPG, and blood pressure) were included as predictors of the target variable, MetS. Interestingly, the proposed CHAID model with the user-specified first split on waist circumference outperformed the CHAID algorithm without first-level split specification and the logistic regression model in both overall accuracy and ability to detect MetS.

The user-specified first split on waist circumference in the decision tree was based on the current literature showing that high waist circumference is the most frequent risk component in people with metabolic syndrome [6] and is highly correlated with diabetes and cardiovascular risks [2, 5]. The IDF guidelines use waist circumference as the first criteria followed by two or more other cardiometabolic abnormalities [16]. However, in these guidelines, the waist circumference criteria would be met for MetS if BMI was greater than 30 kg/m^2 [1]. In the current study, the mean BMI was 29.2 ± 6.8, suggesting that user-specified waist

Table 2: Decision rules for the prediction of the incidence risk of MetS from the CHAID algorithm.

Node number	Level 1	Level 2	Level 3	Level 4	MetS probability
29	94 < waist circumference ≤ 103	TG > 162	HDL ≤ 38	FPG > 94	94.4
11	Waist circumference > 103	HDL ≤ 38	*	*	82.1
17	86 < waist circumference ≤ 94	TG > 138	FPG > 92	*	52.2
12	Waist circumference > 103	38 < HDL ≤ 49	*	*	45.3
28	94 < waist circumference ≤ 103	TG > 162	HDL ≤ 38	FPG ≤ 94	40.0
21	94 < waist circumference ≤ 103	TG > 162	HDL > 38	*	31.8
27	Waist circumference > 103	HDL > 49	FPG > 103	*	26.7
19	94 < waist circumference ≤ 103	TG ≤ 162	FPG > 99	*	18.5
15	86 < waist circumference ≤ 94	TG ≤ 138	FPG > 103	*	16.7
5	Waist circumference ≤ 86	HDL ≤ 38	*	*	4.9
26	Waist circumference > 103	HDL > 49	FPG ≤ 103	*	2.1
6	Waist circumference ≤ 86	HDL > 38	*	*	0.0
14	86 < waist circumference ≤ 94	TG ≤ 138	FPG ≤ 103	*	0.0
16	86 < waist circumference ≤ 94	TG > 138	FPG ≤ 92	*	0.0
18	94 < waist circumference ≤ 103	TG ≤ 162	FPG ≤ 99	*	0.0

*represents not significant. Growing method: exhaustive CHAID; dependent variable: MetS: metabolic syndrome, TG: triglyceride, HDL: high-density lipoprotein cholesterol, and FPG: fasting plasma glucose.

circumference in the decision tree resulted in findings similar to those used by IDF in MetS classification. Two recent studies by Worachartcheewan et al. [20] and Kawada et al. [21] identified the optimal waist circumference cutoff for prediction of MetS. The optimal waist circumference cutoff in the study by Worachartcheewan et al. [20] and Kawada et al. [21] was in the range of 85–88 cm in male and females and 83–85 cm in males, respectively, compared with 86 cm in men and women in the current study. The comparability of these results supports the validity of our findings showing that the CHAID algorithm waist circumference cutoffs could accurately detect MetS.

Central adiposity has been identified as a strong predictor of MetS and a strong contributor to BMI and waist circumference. Després et al. [22] demonstrated a strong correlation between BMI and waist circumference ($r = 0.91$, $P < 0.05$) that is comparable to the current study ($r = 0.93$, data not shown). Furthermore, BMI did not take into consideration the actual body composition, although waist circumference and BMI have been shown to be a strong proxy of visceral adiposity [23]. However, large variances of girth measurements in epidemiological samples weaken the clinical interchangeability between BMI and waist circumference. Waist circumference as compared to BMI might therefore be a more sensitive predictor of MetS, especially in the at-risk young adult population. A waist circumference screening could more readily and easily alert health providers to the increased metabolic risks associated with excessive visceral fat accumulation over other MetS classification criteria that require fasting, blood draws, and analysis. Therefore waist circumference shows promise as an initial predictor in the detection of MetS prior to further testing.

Interestingly, blood pressure did not reach significance to be included in the final model. One possible explanation is that elevations in blood pressure are less prevalent in younger adults [13]. Within our sample, the blood pressure criteria had the lowest prevalence of all the MetS classification criteria for subjects with and without MetS (10.1% and 0.8%, resp.).

4.1. Limitations. The current study was intended as a pilot study meant to explore and test the CHAID algorithm's utility in creating pathways to detect MetS in young adults. Although this model had an overall accuracy of 92.3%, its ability to accurately detect MetS was only 71.8%. The CHAID algorithm requires large sample sizes to operate effectively. Given that the parent and child nodes were set to split at small sizes (20 and 5, resp.) and that there was no validation of the model, the derived pathways for MetS detection from this study are not intended for clinical use. Furthermore, the MetS diagnosis in this analysis was not a clinical diagnosis but was rather determined by the presence of three or more of the NCEP criteria based on their prevalence within the secondary data set. Additionally, this analysis did not account for the use of medications to control blood pressure, lipids, and/or plasma glucose.

The CHAID analysis did not identify any significant differences in MetS based on sex or ethnicity in this sample although previous studies have shown differences in MetS risk based on sex and ethnicity [24, 25]. Considering the limitation of the current study, future investigations warrant utilizing sufficiently large sample sizes, considering the difference in MetS based on the sex and ethnicity and performing model validation.

5. Conclusion

In summary, these preliminary findings suggest that young adults at risk for MetS, who are not routinely screened for fasting blood lipids or FPG, could be identified for further follow-up testing based on their waist circumference. Future research warrants the investigation of other anthropometric

measures, simple point-of-care techniques, and validation of these decision tree methods to create a strong algorithm for predicting and/or the early detection of MetS in young adults. There are no clinically established criteria for premetabolic syndrome. Decision tree methods are promising regarding preliminary MetS detection and can aid in the development of a formal definition of premetabolic syndrome. If established, premetabolic syndrome diagnostic criteria could improve outcomes associated with the development of MetS or could halt the progression of MetS and its relative consequences.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] K. G. M. M. Alberti, R. H. Eckel, S. M. Grundy et al., "Harmonizing the metabolic syndrome: a joint interim statement of the International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity," *Circulation*, vol. 120, no. 16, pp. 1640–1645, 2009.

[2] K. G. M. M. Alberti and P. Zimmet, "The metabolic syndrome—a new worldwide definition," *The Lancet*, vol. 366, no. 9491, pp. 1059–1062, 2005.

[3] Z. T. Bloomgarden, "Consequences of diabetes: cardiovascular disease," *Diabetes Care*, vol. 27, no. 7, pp. 1825–1831, 2004.

[4] J. Ärnlöv, E. Ingelsson, J. Sundström, and L. Lind, "Impact of body mass index and the metabolic syndrome on the risk of cardiovascular disease and death in middle-aged men," *Circulation*, vol. 121, no. 2, pp. 230–236, 2010.

[5] P. Y. Liu, L. M. Hornbuckle, L. B. Panton, J. S. Kim, and J. Z. Ilich, "Evidence for the association between abdominal fat and cardiovascular risk factors in overweight and obese African American women," *Journal of the American College of Nutrition*, vol. 31, no. 2, pp. 126–132, 2012.

[6] A. A. Motala, T. Esterhuizen, F. J. Pirie, and M. A. K. Omar, "The prevalence of metabolic syndrome and determination of the optimal waist circumference cutoff points in a rural South African community," *Diabetes Care*, vol. 34, no. 4, pp. 1032–1037, 2011.

[7] D. M. Boudreau, D. C. Malone, M. A. Raebel et al., "Health care utilization and costs by metabolic syndrome risk factors," *Metabolic Syndrome and Related Disorders*, vol. 7, no. 4, pp. 305–313, 2009.

[8] R. L. Coleman, R. J. Stevens, R. Retnakaran, and R. R. Holman, "Framingham, SCORE, and DECODE risk equations do not provide reliable cardiovascular risk estimates in type 2 diabetes," *Diabetes Care*, vol. 30, no. 5, pp. 1292–1294, 2007.

[9] A. H. Gandomi, M. M. Fridline, and D. A. Roke, "Decision tree approach for soil liquefaction assessment," *The Scientific World Journal*, vol. 2013, Article ID 346285, 8 pages, 2013.

[10] A. Worachartcheewan, C. Nantasenamat, C. Isarankura-Na-Ayudhya, P. Pidetcha, and V. Prachayasittikul, "Identification of metabolic syndrome using decision tree analysis," *Diabetes Research and Clinical Practice*, vol. 90, no. 1, pp. e15–e18, 2010.

[11] R. L. Sacco, M. Khatri, T. Rundek et al., "Improving global vascular risk prediction with behavioral and anthropometric factors. The multiethnic NOMAS (Northern Manhattan Cohort Study)," *Journal of the American College of Cardiology*, vol. 54, no. 24, pp. 2303–2311, 2009.

[12] R. B. D'Agostino Sr., R. S. Vasan, M. J. Pencina et al., "General cardiovascular risk profile for use in primary care: the Framingham heart study," *Circulation*, vol. 117, no. 6, pp. 743–753, 2008.

[13] P. A. Braveman, C. Cubbin, S. Egerter, D. R. Williams, and E. Pamuk, "Socioeconomic disparities in health in the United States: what the patterns tell us," *American Journal of Public Health*, vol. 100, supplement 1, pp. S186–S196, 2010.

[14] A. J. Cameron, P. Z. Zimmet, J. E. Shaw, and K. G. M. M. Alberti, "The metabolic syndrome: in need of a global mission statement," *Diabetic Medicine*, vol. 26, no. 3, pp. 306–309, 2009.

[15] Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics, *2007-2008 National Health and Nutrition Examination Survey: Survey Operations Manuals, Brochures, Consent Documents*, U.S. Department of Health and Human Services, 2013, http://www.cdc.gov/nchs/nhanes/nhanes2007-2008/current_nhanes_07_08.htm.

[16] World Health Organization, *Waist Circumference and Waist-Hip Ratio: Report of a WHO Expert Consultation, Geneva, 8–11 December 2008*, World Health Organization, Geneva, Switzerland, 2008.

[17] G. V. Kass, "An exploratory technique for investigating large quantities for categorical data," *Applied Statistics*, vol. 20, pp. 119–127, 1980.

[18] M. Dehghan and A. T. Merchant, "Is bioelectrical impedance accurate for use in large epidemiological studies?" *Nutrition Journal*, vol. 7, no. 1, article 26, 2008.

[19] V. L. Roger, A. S. Go, D. M. Lloyd-Jones et al., "Heart disease and stroke statistics—2011 update: a report from the american heart association," *Circulation*, vol. 123, no. 4, pp. e18–e209, 2011.

[20] A. Worachartcheewan, P. Dansethakul, C. Nantasenamat, P. Pidetcha, and V. Prachayasittikul, "Determining the optimal cutoff points for waist circumference and body mass index for identification of metabolic abnormalities and metabolic syndrome in urban Thai population," *Diabetes Research and Clinical Practice*, vol. 98, no. 2, pp. e16–e21, 2012.

[21] T. Kawada, T. Otsuka, H. Inagaki et al., "Optimal cut-off levels of body mass index and waist circumference in relation to each component of metabolic syndrome (MetS) and the number of MetS component," *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, vol. 5, no. 1, pp. 25–28, 2011.

[22] J.-P. Després, L. van Gaal, X. Pi-Sunyer, and A. Scheen, "Efficacy and safety of the weight-loss drug rimonabant," *The Lancet*, vol. 371, no. 9612, p. 555, 2008.

[23] K. Lee, S. Lee, Y.-J. Kim, and Y.-J. Kim, "Waist circumference, dual-energy X-ray absortiometrically measured abdominal adiposity, and computed tomographically derived intra-abdominal fat area on detecting metabolic risk factors in obese women," *Nutrition*, vol. 24, no. 7-8, pp. 625–631, 2008.

[24] P. Hari, K. Nerusu, V. Veeranna et al., "A gender-stratified comparative analysis of various definitions of metabolic syndrome

and cardiovascular risk in a multiethnic U.S. population," *Metabolic Syndrome and Related Disorders*, vol. 10, no. 1, pp. 47–55, 2012.

[25] S. E. Walker, M. J. Gurka, M. N. Oliver, D. W. Johns, and M. D. DeBoer, "Racial/ethnic discrepancies in the metabolic syndrome begin in childhood and persist after adjustment for environmental factors," *Nutrition, Metabolism and Cardiovascular Diseases*, vol. 22, no. 2, pp. 141–148, 2012.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 746979, 8 pages
http://dx.doi.org/10.1155/2014/746979

Research Article

Establishing Reliable miRNA-Cancer Association Network Based on Text-Mining Method

Lun Li,[1,2] Xingchi Hu,[1,2] Zhaowan Yang,[1,2] Zhenyu Jia,[3,4,5] Ming Fang,[1,2] Libin Zhang,[1,2] and Yanhong Zhou[1,2]

[1] *Hubei Bioinformatics and Molecular Imaging Key Laboratory, Huazhong University of Science and Technology, Wuhan 430074, China*
[2] *Biomedical Engineering Department, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China*
[3] *Department of Statistics, The University of Akron, Akron, OH 44325, USA*
[4] *Department of Family & Community Medicine, Northeast Ohio Medical University, Rootstown, OH 44272, USA*
[5] *Guizhou Provincial Key Laboratory of Computational Nano-Material Science, Guizhou Normal College, Guiyang 550018, China*

Correspondence should be addressed to Libin Zhang; libinzhang@mail.hust.edu.cn and Yanhong Zhou; yhzhou@mail.hust.edu.cn

Received 24 January 2014; Revised 5 March 2014; Accepted 6 March 2014; Published 10 April 2014

Academic Editor: Xiao-Qin Xia

Associating microRNAs (miRNAs) with cancers is an important step of understanding the mechanisms of cancer pathogenesis and finding novel biomarkers for cancer therapies. In this study, we constructed a miRNA-cancer association network (miCancerna) based on more than 1,000 miRNA-cancer associations detected from millions of abstracts with the text-mining method, including 226 miRNA families and 20 common cancers. We further prioritized cancer-related miRNAs at the network level with the random-walk algorithm, achieving a relatively higher performance than previous miRNA disease networks. Finally, we examined the top 5 candidate miRNAs for each kind of cancer and found that 71% of them are confirmed experimentally. miCancerna would be an alternative resource for the cancer-related miRNA identification.

1. Introduction

MicroRNAs (miRNAs) are a large class of small noncoding RNAs [1] known to be functionally involved in a wide range of biological processes including embryo development, cell growth, differentiation, apoptosis, and proliferation [2–5]. Recently, it has been found that miRNAs play important roles in human tumor genesis and many of them have also been applied as novel biomarkers for cancer therapies [6–11], which attracts more and more efforts in revealing the complex associations between miRNAs and cancers. However, the existing literature usually focused on the relationship between several miRNAs and a specific cancer, leaving the comprehensive miRNA-cancer network unrevealed. Therefore, fully uncovering the associations between miRNAs and cancers would be extremely interesting and valuable for identifying cancer-related miRNA and understanding the mechanisms behind.

To this aim, the manually collected miRNA-disease association databases HMDD [12] and miR2Disease [13] have been established. At present, these manually created miRNA-disease networks have been used to predict disease-related miRNAs [14–16] and achieved relatively high accuracies, opening opportunity of prioritizing miRNAs with bioinformatics methods.

However, thousands of papers on miRNA and cancer researches are published each year, making it difficult to manually check papers. On the other hand, automatic text-mining methods are needed to extract reliable miRNA-disease associations [17] from the increasing database.

In this paper, we collected 1,018 associations between 226 miRNA families and 20 common cancers by extracting from more than 7.1 million publications with an automatic text-mining method. All these relationships have been recorded in a database named miCancerna, which can be freely assessed at http://micancerna.appspot.com/. We further constructed a miRNA-cancer general view on top 5% significant associations for visualizing the roles of miRNAs in different cancers and prioritized the cancer-related miRNAs using the random

walk with restart algorithm (RWRA) [14] on miRNA-cancer network built on the data in miCancerna. By analyzing the top 5 associated miRNAs of 20 cancers according to Fisher's exact tests, we found experimental evidence for 71% of these miRNA-cancer relationships, and the rest might be candidate cancer-related miRNAs for further experimental validation. The constructed miRNA-cancer network would be extremely valuable for comprehensively understanding the mechanisms of cancers and identifying cancer-related miRNA genes.

2. Materials and Methods

2.1. Collecting Resource Literature. We collected the abstracts from NCBI's MEDLINE database as our target literature resource. MEDLINE is a comprehensive database containing the abstracts of millions of articles in biomedical area. Since a large number of papers are not fully accessible in the PubMed database, we only consider the abstracts for the papers, which are always available.

In 2000, Reinhart et al. [18] identified the second miRNA, and thereafter researchers began to pay attention to the importance of miRNAs. Therefore, we mainly focus on the papers that have been published in 2000 and after. In total, 7,207,066 abstracts were retrieved and then screened using keywords, such as "Humans" or "Animals," within the PubMed search for eliminating plant and virus miRNAs in the following text-mining analysis. This filtration yielded 5,606,308 paper abstracts.

Currently, the 20 most common cancers reported by National Cancer Institute (http://www.cancer.gov/) are considered in our study, including leukemia, lung cancer bladder cancer, brain cancer, breast cancer, cervix cancer, colorectal cancer, esophageal cancer, kidney cancer, liver cancer, melanoma, myeloma, non-Hodgkin lymphoma, oral cancer, ovarian cancer, pancreatic cancer, prostate cancer, stomach cancer, thyroid cancer, and uterine cancer. The abstracts are individually marked with cancer types by the following steps: first, we mapped each cancer type to its corresponding MeSH (medical subject headings) term(s), the U.S. National Library of Medicine's controlled vocabulary that are manually assigned for articles archived in MEDLINE describing their subject matters, and then compiled a list of standard names of each type of cancer. Subsequently, we searched each article abstract for the MeSH annotations. The abstracts with MeSH terms in our cancers name list are marked with the corresponding cancer and selected for the following text-mining processing.

2.2. Establishing miRNA-Cancer Networks by Text-Mining Method. With the selected abstracts, we firstly established relationships between miRNAs and cancers by a text-mining method. The associations between miRNAs and cancers were estimated based on the cooccurrence assumption, which is the fundamental assumption in the field of text-mining and can be used to infer whether two terms are associated or not. In our case, if a particular miRNA appears in the abstracts marked by a specific cancer frequently, we can reasonably assume that they cooccurred and tend to be related. To establish the associations between miRNAs and cancers, we detect the appearance of miRNAs in the abstracts marked by cancer types. In this study, the regular expression was applied to match miRNA names against the texts with the following steps. (1) miRNAs (such as "miR-1" and "miR-2") were firstly extracted from the abstracts with the nomenclature of a "miR" prefix accompanied by a unique identifying number [19]. (2) Following the conventions, a prefixed species/state identifier can be added (e.g., "hsa-miR-1" in *Homo sapiens* and "pre-miR-1" for a precursor) and additional suffixes can be given to indicate loci or variant (e.g., "miR-1a-1") [20]. (3) The regular expression was also designed for the variants of some miRNAs, such as "lin-4" and "let-7." (4) Abbreviations for more than one miRNA are also recognized by the regular expression, for example, "miR-221/222" and "miR-15 & -16."

The significance levels of the associations of the miRNAs and the cancers extracted from the marked abstracts were estimated by one-sided Fisher's exact tests [21]. For a pair of the miRNA M and the cancer C, the P value of Fisher's exact test is calculated based on hypergeometric distribution, as follows: $P = \binom{a+b}{a}\binom{c+d}{c} / \binom{n}{a+c} = (a+b)!(c+d)!(a+c)!(c+d)!/(a!b!c!d!n!)$, where n is denoted as the total number of papers included in the text-mining analysis, a stands for the number of papers with both the miRNA M and the cancer C in the abstracts, b and c represent, respectively, the number of abstracts containing one termand excluding the other, and d is the number of papers with neither of the terms. The top 5% miRNA-cancer associations with the minimum P value are considered as significant and were used to generate the general view for miRNA-cancer network. The miRNA-cancer network is a bipartite network composed by miRNA nodes and cancer nodes. Each edge in miCancerna connects a miRNA and one of its corresponding cancers.

2.3. Text-Mining Quality Check. We first queried PubMed with "MIR or MIRN or MIRNA or MICRORNA" and randomly picked up 100 MEDLINE abstracts with at least one miRNA identifier from the querying result as our evaluating data. We then investigated the reliability of detecting miRNAs in texts using the F-measure, which is the harmonic mean of two other measures, recall and precision, as follows:

$$\begin{aligned} \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ F\text{-measure} &= \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}, \end{aligned} \tag{1}$$

where TP, FP, and FN are the number of true positives, false positives, and false negatives, respectively.

2.4. Random Walk with Restart Method. Based on the network constructed by the data from miCancerna, a random walk with restart (RWRA) method is applied to prioritize cancer-related miRNAs.

RWRA is one of the random walk models widely used in disease gene discovery [22]. It simulates a random walker's

moves in a given network and the walker moves from a current node to a direct neighboring node or restart with a training node with the probability (α). The movement given out by RWRA is defined as follows:

$$P_{t+1} = (1 - \alpha)\, MP_t + \alpha P_0, \tag{2}$$

where M is a column-normalized adjacency matrix representing the given network. In this case, each nonzero node in M stands for a certain association between a miRNA and a cancer, and these nodes are taken as seeds. P_t is a vector representing the probabilities of the walker at each node at time t, and P_0 is the initial probability vector in which training nodes are equally assigned $1/N$ (N is the number of seeds) while others are 0. The process is iterated until P reaches a stable status when the difference between P_{t+1} and P_t (measured by $L1$ norm) is less than a threshold value (10^{-6} in this study). The stable probability is defined as P_∞. The candidate nodes are then ranked in descending order according to P_∞.

2.5. Leave-One-Out Cross-Validation. The performance of cancer-related miRNA prioritization by random walk with restart algorithm through miCancerna could be evaluated by calculating the area under the ROC through the leave-one-out cross-validation. For each training node, we took it as a candidate node and randomly picked 20 miRNAs not belonging to the same cancer as testing nodes and then prioritized them as above. For each threshold, the sensitivity (SN) and specificity (SP) are defined as follows:

$$\begin{aligned} \mathrm{SN} &= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \\ \mathrm{SP} &= \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}}, \end{aligned} \tag{3}$$

where TP (true positive) is the number of training nodes with rank above the threshold, FN (false negative) is the number of training nodes with rank under the threshold, TN (true negative) is the number of testing nodes with rank under the threshold, and FP (false positive) is the number of test nodes with rank above the threshold. The ROC curve shows the relationship between SN and 1-SP, and the AUC means the area under the ROC curve.

3. Result and Discussion

3.1. Online Resource for miRNA-Cancer Network. In the first release, miCancerna records 1,018 associations between 226 miRNA families and 20 common cancers extracted from 7.2 million papers. Now all the data that miCancerna refers to can be freely assessed at http://micancerna.appspot.com/, including the associations, the supporting papers, and significant levels for each association. miCancerna will be updated periodically.

To check the text-mining quality, we randomly picked up 100 MEDLINE abstracts that contained at least one miRNA identifier from the search results by querying MEDLINE with "MIR or MIRN or MIRNA or MICRORNA." A total of 739 miRNA identifiers were manually recognized in the texts of evaluating data, while our regular expression correctly matched 735 of them (true positive, TP), miscalled 2 (false positive, FP), and missed 4 (false negative, FN). So the miRNA annotation gained recall of 0.9946, precision of 0.9973, and F-measure of 0.9959, which demonstrated a fairly high reliability of our regular expression.

Table 1: Top 10 associates between miRNAs and cancers.

miRNA	Cancer	Papers	P value
miR-15	Leukaemia	35	6.804×10^{-43}
miR-16	Leukaemia	33	5.028×10^{-36}
miR-122	Liver cancer	22	9.742×10^{-26}
miR-181	Leukaemia	23	3.142×10^{-25}
miR-155	Non-Hodgkin lymphoma	22	7.393×10^{-22}
Let-7	Lung cancer	34	1.110×10^{-19}
miR-223	Leukaemia	16	1.987×10^{-18}
miR-17	Non-Hodgkin lymphoma	19	3.772×10^{-18}
miR-21	Breast cancer	31	1.659×10^{-16}
miR-221	Thyroid cancer	11	1.607×10^{-14}

According to these comparison results, we concluded that miCancerna is a high-quality resource of miRNA-cancer associations.

3.2. miRNA-Cancer Network Visualization. To reveal the roles of miRNA in different cancers, we constructed a bipartite network with the top 5% associations based on Fisher's exact test P values in miCancerna, consisting of 40 miRNA families and 13 types of cancers (Figure 1). In this bipartite network, miRNAs are only connected to cancers and cancers are only connected to miRNAs. The miRNA-cancer network was visualized with Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/). It is interesting to find that almost all these cancers (except the stomach cancer) can be connected via miRNAs, which indicated that different cancers might share common pathogenic components regulated by these interconnected miRNAs, while stomach cancer may be different with others.

As shown in Figure 1, miRNAs may have different involvements in cancers. Some miRNAs are specifically associated with a specific cancer. For example, miR-15 and miR-16 are tendentiously related to leukemia, and miR-122 is almost exclusively associated with liver cancer. These miRNAs may be used as biomarker candidates for diagnosis and efficacy of therapies for corresponding cancers. By contrast, some miRNAs tend to be associated with various cancers. One example is miR-21, which is shown to significantly associate with breast cancer, colorectal cancer, liver cancer, and pancreatic cancer, indicating that target genes of miR-21 might play critical roles in tumor formation.

It is interesting that four miRNA-cancer associations in top 10 (Table 1) are miRNA-leukemia associations, and 28.6% (12) of significant associations were related to leukemia, which makes leukemia the most miRNA-related cancer. Similarly, 8 (19.0%) miRNA families were related to breast cancer in significant miRNA-cancer associations. Furthermore, we

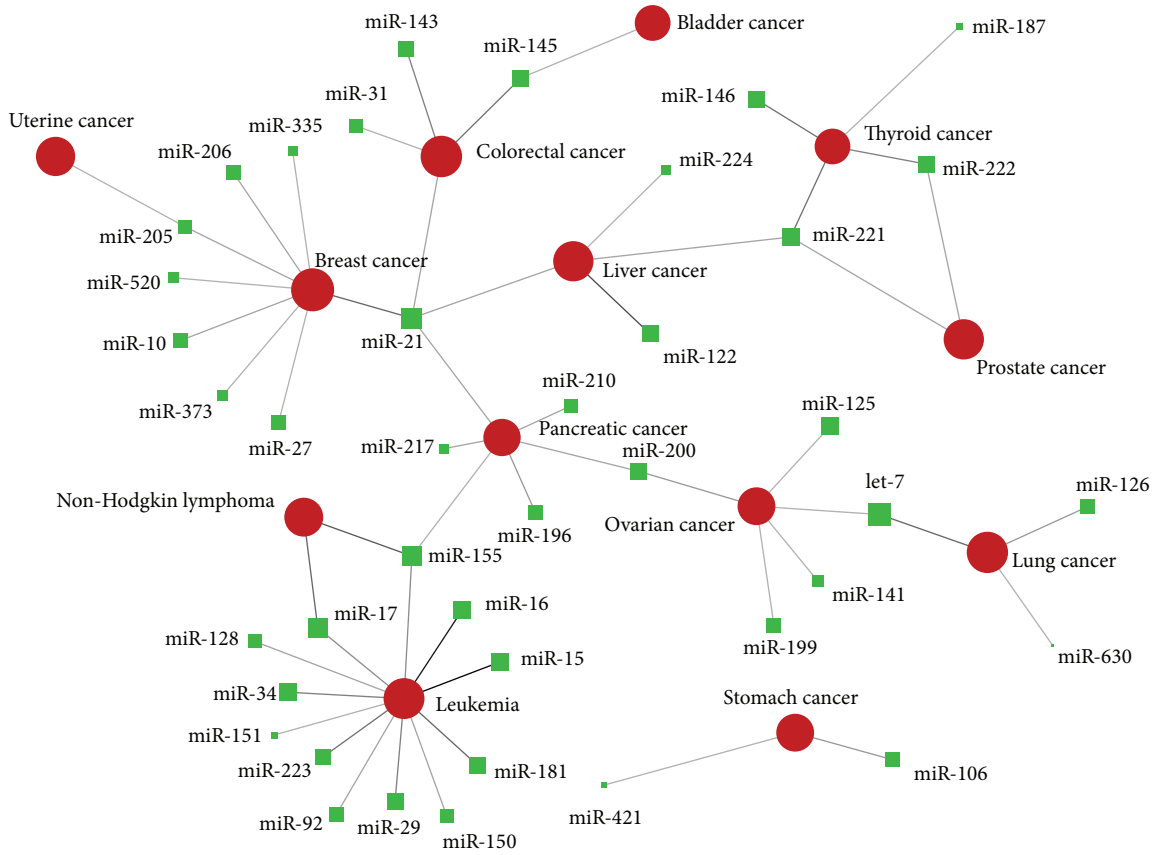


Figure 1: Network illustrated significant associations of miRNAs and cancers. Red circles and green squares represent cancers and miRNAs, respectively, with different sizes according to the number of corresponding annotated papers (logarithmic). Each link represents a miRNA-cancer association with colour and width according to the strength of relationship.

found that miR-21 is the most cancer-related miRNA, which is associated with 4 (30.77%) different cancers in significant associations (breast cancer, pancreatic cancer, liver cancer, and colorectal cancers), indicating that miR-21 may be involved in an important pathway in cancer formation.

3.3. Prioritization of Cancer-Related miRNAs. We applied RWRA on the network established by miCancerna to prioritize candidate cancer-related miRNAs, and the performance is evaluated by leave-one-out cross-validation. With a restart probability alpha of 0.9, the AUC of ROC curve can reach 0.798 (Figure 2), while the AUC of 1 stands for the perfect performance and AUC of 0.5 indicates the random performance. The performances with different restart probabilities are showed in Table 2. The AUC improves as alpha increases, but the variation is small. To rule out the possibility that the performance of miCancerna is achieved by chance, a permutation test with 300 runs was performed. For each run, the seeds are randomly selected from the candidate nodes. The average AUC of random permutations obtained by leave-one-out cross validation is 0.513, and the distribution of the random permutation AUCs is shown in Figure 3. It is obvious that there is significant difference between the AUC achieved by miCancerna and the random permutations, which supports

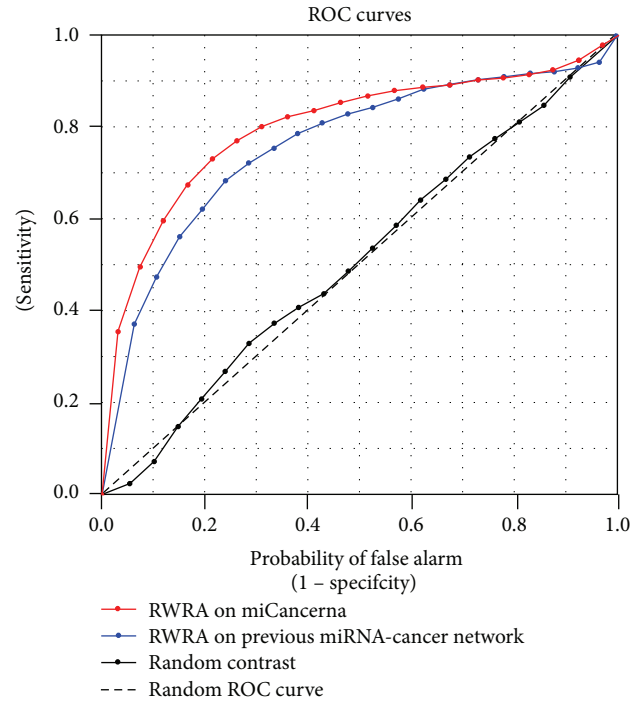


Figure 2: ROC curves for RWRA on miCancerna and previous miRNA-cancer network.

Table 2: AUC value under different alpha.

Alpha	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
AUC	0.7952	0.7973	0.7974	0.7978	0.7981	0.7981	0.7983	0.7983	0.7984

Table 3: Top 5 potential miRNAs of 20 cancers.

Bladder cancer		Brain cancer		Breast cancer		Cervix cancer	
miRNAs	Confirm	miRNAs	Confirm	miRNAs	Confirm	miRNAs	Confirm
miR-15	Null	let-7	Ref. [25]	miR-143	dbDEMC	let-7	Null
miR-34	Ref. [26]	miR-145	Ref. [27]	miR-223	dbDEMC	miR-221	Null
miR-16	Ref. [26]	miR-16	Ref. [28]	miR-203	dbDEMC	miR-17	Ref. [29]
miR-146	Ref. [30]	miR-155	Ref. [31]	miR-194	dbDEMC	miR-125	Null
miR-155	Ref. [30]	miR-143	Ref. [28]	miR-100	dbDEMC	miR-222	Null
Colorectal cancer		Esophageal cancer		Kidney cancer		Leukemia	
miRNAs	Confirm	miRNAs	Confirm	miRNAs	Confirm	miRNAs	Confirm
miR-221	dbDEMC	miR-17	dbDEMC	miR-125	dbDEMC	miR-200	Ref. [32]
miR-146	dbDEMC	miR-222	dbDEMC	miR-222	dbDEMC	miR-205	Null
miR-29	dbDEMC	miR-15	dbDEMC	miR-146	dbDEMC	miR-193	Null
miR-199	dbDEMC	miR-125	dbDEMC	miR-16	dbDEMC	miR-9	Ref. [33]
miR-193	Null	miR-200	dbDEMC	miR-143	dbDEMC	miR-31	Ref. [34]
Liver cancer		Lung cancer		Melanoma		Myeloma	
miRNAs	Confirm	miRNAs	Confirm	miRNAs	Confirm	miRNAs	Confirm
miR-205	Null	miR-23	dbDEMC	miR-21	Ref. [35]	miR-145	Null
miR-27	dbDEMC	miR-148	dbDEMC	miR-145	Ref. [36]	miR-200	Null
miR-124	Ref. [37]	miR-27	dbDEMC	miR-26	Null	miR-221	Ref. [38]
miR-520	dbDEMC	miR-203	dbDEMC	miR-143	Ref. [36]	miR-34	Null
miR-203	Ref. [39]	miR-520	dbDEMC	miR-126	Ref. [35]	miR-205	Null
Non-Hodgkin lymphoma		Oral cancer		Ovarian cancer		Pancreatic cancer	
miRNAs	Confirm	miRNAs	Confirm	miRNAs	Confirm	miRNAs	Confirm
miR-200	dbDEMC	miR-15	Null	miR-26	Null	miR-16	Ref. [40]
miR-205	dbDEMC	miR-205	Ref. [41]	miR-181	Null	miR-125	Ref. [42]
miR-126	dbDEMC	miR-10	Ref. [43]	miR-143	Ref. [44]	miR-26	Null
miR-224	dbDEMC	miR-182	Null	miR-10	Null	miR-126	Ref. [45]
miR-23	dbDEMC	miR-20	Null	miR-23	Null	miR-181	Ref. [40]
Prostate cancer		Stomach cancer		Thyroid cancer		Uterine cancer	
miRNAs	Confirm	miRNAs	Confirm	miRNAs	Confirm	miRNAs	Confirm
miR-155	dbDEMC	miR-155	Ref. [46]	miR-15	Null	miR-17	dbDEMC
miR-29	Null	miR-29	Null	miR-34	Null	miR-222	dbDEMC
miR-30	dbDEMC	miR-30	Null	miR-145	Ref. [47]	miR-224	dbDEMC
miR-10	dbDEMC	miR-10	Ref. [48]	miR-16	Null	miR-30	dbDEMC
miR-199	dbDEMC	miR-199	Null	miR-205	Ref. [49]	miR-106	dbDEMC

"Null" means we did not find experimental evidence.

that the miCancerna reveals the real involvement of miRNAs in cancer biology.

The top 5 potential miRNAs of each cancer are presented in Table 3, among which 71% have been evaluated by experimental evidence in dbDEMC [23] or literatures published after miCancerna. The performance of cancer-related miRNA prioritization demonstrates the reliability of miCancerna. Moreover, the top predicted miRNAs may be the potential cancer-related miRNAs for further study.

3.4. Comparison with Similar Databases. We made comparisons with similar database or networks. First we compared the data involved in miCancerna and the manual checking database miR2Disease on the number of evidence papers. For most cancers, miCancerna provides much more evidence papers than miR2Disease (Table 4). Second, we compared the prediction performance of RWRA on miCancerna with the miRNA-cancer network used in RWRMDA [14], which was built based on HMDD, a manual database. The ROC curves

Table 4: Number of evidence papers comparing with miR2Diease.

Cancer types	miCancerna	miR2Disease	Increase
Bladder cancer	14	11	27.27%
Brain cancer	35	3	1067%
Breast cancer	137	58	136.2%
Cervix cancer	11	4	175%
Colorectal cancer	81	39	107.7%
Esophageal cancer	16	7	128.6%
Kidney cancer	14	4	250.0%
Leukemia	146	45	224.4%
Liver cancer	99	39	153.8%
Lung cancer	112	37	202.7%
Melanoma	21	9	133.3%
Myeloma	9	3	200.0%
Non-Hodgkin lymphoma	62	13	376.9%
Oral cancer	19	0	—
Ovarian cancer	47	18	161.1%
Pancreatic cancer	47	16	193.8%
Prostate cancer	61	19	221.1%
Stomach cancer	48	16	200.0%
Thyroid cancer	21	9	133.3%
Uterine cancer	28	5	460.0%

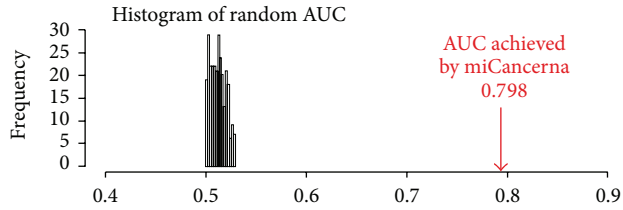


Figure 3: Distribution of random AUC for miCancerna.

for both networks are showed in Figure 2. According to the result of leave-one-out cross-validation, the network used in RWRMDA achieved AUC of 0.763, which is lower than 0.797 achieved by miCancerna.

These results indicate that miCancerna provides an alternative resource of miRNA-cancer associations.

4. Conclusion

In this study, we constructed a reliable miRNA-cancer network based on text-mining method, which is stored in the database miCancerna. In current release, there are 1,018 associations between 226 miRNA families and 20 common cancers. According to our test result, the miCancerna provides a reliable and comprehensive resource of miRNA-cancer associations, which can be further used in the identification of cancer-related miRNAs.

For future development, we plan to consider more types of cancers, add regulation information to the miRNA-cancer associations, and integrate miCancerna into other related databases, such as MISIM [24], the human miRNA functional similarity and functional network.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contribution

Lun Li, Xingchi Hu, and Zhaowan Yang contributed equally to this work.

Acknowledgments

This study was supported by the National Natural Science Foundation of China [30971642], the Natural Science Foundation of Hubei Province of China [2009CDA161], and National Natural Science Foundation of China [31000570]. The authors thank Professor Anyuan Guo for his advice and Yanhua Jiang for the help in drafting and revising the paper.

References

[1] D. P. Bartel, "MicroRNAs: genomics, biogenesis, mechanism, and function," *Cell*, vol. 116, no. 2, pp. 281–297, 2004.

[2] D. P. Bartel, "MicroRNAs: genomics, biogenesis, mechanism, and function," *Cell*, vol. 116, no. 2, pp. 281–297, 2004.

[3] A. M. Cheng, M. W. Byrom, J. Shelton, and L. P. Ford, "Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis," *Nucleic Acids Research*, vol. 33, no. 4, pp. 1290–1297, 2005.

[4] B. D. Harfe, "MicroRNAs in vertebrate development," *Current Opinion in Genetics and Development*, vol. 15, no. 4, pp. 410–415, 2005.

[5] E. Wienholds, W. P. Kloosterman, E. Miska et al., "Cell biology: microRNA expression in zebrafish embryonic development," *Science*, vol. 309, no. 5732, pp. 310–311, 2005.

[6] A. A. Shah, P. Leidinger, N. Blin, and E. Meese, "miRNA: small molecules as potential novel biomarkers in cancer," *Current Medicinal Chemistry*, vol. 17, no. 36, pp. 4427–4432, 2010.

[7] S. Huang and X. He, "The role of microRNAs in liver cancer progression," *British Journal of Cancer*, vol. 104, no. 2, pp. 235–240, 2011.

[8] V. A. Krutovskikh and Z. Herceg, "Oncogenic microRNAs (OncomiRs) as a new class of cancer biomarkers," *BioEssays*, vol. 32, no. 10, pp. 894–904, 2010.

[9] J. C. Brase, D. Wuttig, R. Kuner, and H. Sültmann, "Serum microRNAs as non-invasive biomarkers for cancer," *Molecular Cancer*, vol. 9, article 306, 2010.

[10] T. A. Farazi, J. I. Spitzer, P. Morozov, and T. Tuschl, "MiRNAs in human cancer," *Journal of Pathology*, vol. 223, no. 2, pp. 102–115, 2011.

[11] H. Tazawa, S. Kagawa, and T. Fujiwara, "MicroRNAs as potential target gene in cancer gene therapy of gastrointestinal tumors," *Expert Opinion on Biological Therapy*, vol. 11, no. 2, pp. 145–155, 2011.

[12] M. Lu, Q. Zhang, M. Deng et al., "An analysis of human microRNA and disease associations," *PLoS ONE*, vol. 3, no. 10, Article ID e3420, 2008.

[13] Q. Jiang, Y. Wang, Y. Hao et al., "miR2Disease: a manually curated database for microRNA deregulation in human disease," *Nucleic Acids Research*, vol. 37, no. 1, pp. D98–D104, 2009.

[14] X. Chen, M. X. Liu, and G. Y. Yan, "RWRMDA: predicting novel human microRNA-disease associations," *Molecular BioSystems*, vol. 8, no. 10, pp. 2792–2798, 2012.

[15] Q. Jiang, Y. Hao, G. Wang et al., "Prioritization of disease microRNAs through a human phenome-microRNAome network," *BMC Systems Biology*, vol. 4, no. 1, article S2, 2010.

[16] H. Chen and Z. Zhang, "Prediction of associations between OMIM diseases and microRNAs by random walk on OMIM disease similarity network," *The Scientific World Journal*, vol. 2013, Article ID 204658, 6 pages, 2013.

[17] Y. A. Lussier, W. M. Stadler, and J. L. Chen, "Advantages of genomic complexity: bioinformatics opportunities in microRNA cancer signatures," *Journal of the American Medical Informatics Association*, vol. 19, no. 2, pp. 156–160, 2012.

[18] B. J. Reinhart, F. J. Slack, M. Basson et al., "The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*," *Nature*, vol. 403, no. 6772, pp. 901–906, 2000.

[19] V. Ambros, B. Bartel, D. P. Bartel et al., "A uniform system for microRNA annotation," *RNA*, vol. 9, no. 3, pp. 277–279, 2003.

[20] S. Griffiths-Jones, H. K. Saini, S. Van Dongen, and A. J. Enright, "miRBase: tools for microRNA genomics," *Nucleic Acids Research*, vol. 36, no. 1, pp. D154–D158, 2008.

[21] R. A. Fisher, "On the mathematical foundations of theoretical statistics," *Philosophical Transactions of the Royal Society of London Series A*, vol. 222, pp. 309–368, 1922.

[22] D.-H. Le and Y.-K. Kwon, "GPEC: a cytoscape plug-in for random walk-based gene prioritization and biomedical evidence collection," *Computational Biology and Chemistry*, vol. 37, pp. 17–23, 2012.

[23] Z. Yang, F. Ren, C. Liu et al., "DbDEMC: a database of differentially expressed miRNAs in human cancers," *BMC Genomics*, vol. 11, no. 4, article S5, 2010.

[24] D. Wang, J. Wang, M. Lu, F. Song, and Q. Cui, "Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases," *Bioinformatics*, vol. 26, no. 13, Article ID btq241, pp. 1644–1650, 2010.

[25] K.-H. Ling, P. J. Brautigan, C. N. Hahn et al., "Deep sequencing analysis of the developing mouse brain reveals a novel microRNA," *BMC Genomics*, vol. 12, article 176, 2011.

[26] R. L. Vinall, M. S. Kent, and D. R. W. White, "Expression of microRNAs in urinary bladder samples obtained from dogs with grossly normal bladders, inflammatory bladder disease, or transitional cell carcinoma," *American Journal of Veterinary Research*, vol. 73, no. 10, pp. 1626–1633, 2012.

[27] M. C. Speranza, V. Frattini, F. Pisati, D. Kapetis, P. Porrati et al., "NEDD9, a novel target of miR-145, increases the invasiveness of glioblastoma," *Oncotarget*, vol. 3, pp. 723–734, 2012.

[28] M. L. Campanini, L. M. Colli, B. M. C. Paixao et al., "CTNNB1 gene mutations, pituitary transcription factors, and MicroRNA expression involvement in the pathogenesis of adamantinomatous craniopharyngiomas," *Hormones and Cancer*, vol. 1, no. 4, pp. 187–196, 2010.

[29] H.-W. Kang, F. Wang, Q. Wei et al., "miR-20a promotes migration and invasion by regulating TNKS2 in human cervical cancer cells," *FEBS Letters*, vol. 586, no. 6, pp. 897–904, 2012.

[30] G. Wang, E. S. Chan, B. C. Kwan, P. K. Li, S. K. Yip et al., "Expression of microRNAs in the urine of patients with bladder cancer," *Clinical Genitourinary Cancer*, vol. 10, no. 2, pp. 106–113, 2012.

[31] P. Poltronieri, P. I. D'Urso, V. Mezzolla, and O. F. D'Urso, "Potential of anti-cancer therapy based on anti-miR-155 oligonucleotides in glioma and brain tumours," *Chemical Biology and Drug Design*, vol. 81, no. 1, pp. 79–84, 2013.

[32] J. Rosati, F. Spallotta, S. Nanni et al., "Smad-interacting protein-1 and microRNA 200 family define a nitric oxide-dependent molecular circuitry involved in embryonic stem cell mesendoderm differentiation," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 31, no. 4, pp. 898–907, 2011.

[33] D. Chen, Y. Sun, Y. Wei, P. Zhang, A. H. Rezaeian et al., "LIFR is a breast cancer metastasis suppressor upstream of the Hippo-YAP pathway and a prognostic marker," *Nature Medicine*, vol. 18, no. 10, pp. 1511–1517, 2012.

[34] O. H. Rokah, G. Granot, A. Ovcharenko, S. Modai, M. Pasmanik-Chor et al., "Downregulation of miR-31, miR-155, and miR-564 in chronic myeloid leukemia cells," *PLoS ONE*, vol. 7, no. 4, Article ID e35501, 2012.

[35] M. Bhaskaran, D. Xi, Y. Wang, C. Huang, T. Narasaraju et al., "Identification of microRNAs changed in the neonatal lungs in response to hyperoxia exposure," *Physiological Genomics*, vol. 44, no. 20, pp. 970–980, 2012.

[36] M. Nugent, N. Miller, and M. J. Kerin, "Circulating *miR-34a* levels are reduced in colorectal cancer," *Journal of Surgical Oncology*, vol. 106, no. 8, pp. 947–952, 2012.

[37] R. F. Schwabe and T. C. Wang, "Targeting liver cancer: first steps toward a miRacle?" *Cancer Cell*, vol. 20, no. 6, pp. 698–699, 2011.

[38] J.-J. Huang, J. Yu, J.-Y. Li, Y.-T. Liu, and R.-Q. Zhong, "Circulating microRNA expression is associated with genetic subtype and survival of multiple myeloma," *Medical Oncology*, vol. 29, no. 4, pp. 2402–2408, 2012.

[39] M. Furuta, K.-I. Kozaki, S. Tanaka, S. Arii, I. Imoto, and J. Inazawa, "*miR-124* and *miR-203* are epigenetically silenced tumor-suppressive microRNAs in hepatocellular carcinoma," *Carcinogenesis*, vol. 31, no. 5, pp. 766–776, 2009.

[40] Y. Ren, J. Gao, J. Q. Liu, X. W. Wang, J. J. Gu et al., "Differential signature of fecal microRNAs in patients with pancreatic cancer," *Molecular Medicine Reports*, vol. 6, pp. 201–209, 2012.

[41] E. D. Wiklund, S. Gao, T. Hulf et al., "MicroRNA alterations and associated aberrant DNA methylation patterns across multiple sample types in oral squamous cell carcinoma," *PLoS ONE*, vol. 6, no. 11, Article ID e27840, 2011.

[42] Y. Zhao, P. Xie, and H. Fan, "Genomic profiling of MicroRNAs and proteomics reveals an early molecular alteration associated with tumorigenesis induced by MC-LR in mice," *Environmental Science and Technology*, vol. 46, no. 1, pp. 34–41, 2012.

[43] P. Murino, M. Mammucari, D. Borrelli et al., "Role of immediate-release morphine (MIR) in the treatment of predictable pain in radiotherapy," *Journal of Pain and Palliative Care Pharmacotherapy*, vol. 25, no. 2, pp. 121–124, 2011.

[44] S. Marchini, D. Cavalieri, R. Fruscio et al., "Association between miR-200c and the survival of patients with stage I epithelial ovarian cancer: a retrospective study of two independent tumour tissue collections," *The Lancet Oncology*, vol. 12, no. 3, pp. 273–285, 2011.

[45] L. R. Jiao, A. E. Frampton, J. Jacob et al., "Micrornas targeting oncogenes are down-regulated in pancreatic malignant transformation from benign tumors," *PLoS ONE*, vol. 7, no. 2, Article ID e32068, 2012.

[46] C. L. Li, H. Nie, M. Wang, L. P. Su, J. F. Li et al., "microRNA-155 is downregulated in gastric cancer cells and involved in cell metastasis," *Oncology Reports*, vol. 27, no. 6, pp. 1960–1966, 2012.

[47] C. H. Zhou, S. F. Yang, and P. Q. Li, "Human Lung Cancer cell line SPC-A1 contains cells with characteristics of cancer stem cells," *Neoplasma*, vol. 59, no. 6, pp. 685–692, 2012.

[48] Z. Liu, J. Zhu, H. Cao, H. Ren, and X. Fang, "miR-10b promotes cell invasion through RhoC-AKT signaling pathway by targeting HOXD10 in gastric cancer," *International Journal of Oncology*, vol. 40, no. 5, pp. 1553–1560, 2012.

[49] J. A. Bishop, H. Benjamin, H. Cholakh, A. Chajut, D. P. Clark, and W. H. Westra, "Accurate classification of non-small cell lung carcinoma using a novel microRNA-based approach," *Clinical Cancer Research*, vol. 16, no. 2, pp. 610–619, 2010.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 643457, 5 pages
http://dx.doi.org/10.1155/2014/643457

Research Article

Weighted Lin-Wang Tests for Crossing Hazards

James A. Koziol[1] and Zhenyu Jia[2,3,4]

[1] *College of Health, Human Services and Science, Ashford University, San Diego, CA 92128, USA*
[2] *Department of Statistics, University of Akron, Akron, OH 44325, USA*
[3] *Department of Family and Community Medicine, Northeast Ohio Medical University, Rootstown, OH 44272, USA*
[4] *Guizhou Provincial Key Laboratory of Computational Nano-Material Science, Guizhou Normal College, Guiyang 550018, China*

Correspondence should be addressed to Zhenyu Jia; zjia@uakron.edu

Received 2 January 2014; Accepted 17 February 2014; Published 30 March 2014

Academic Editor: Xiao-Qin Xia

Lin and Wang have introduced a quadratic version of the logrank test, appropriate for situations in which the underlying survival distributions may cross. In this note, we generalize the Lin-Wang procedure to incorporate weights and investigate the performance of Lin and Wang's test and weighted versions in various scenarios. We find that weighting does increase statistical power in certain situations; however, none of the procedures was dominant under every scenario.

1. Introduction

Lin and Wang [1] have recently introduced an ingenious modification of the two-sample logrank statistic, appropriate for crossing hazards alternatives. Through a simulation study, they demonstrated that their modified test had greater power than the commonly used logrank and Wilcoxon tests for detecting differences between crossing survival curves. In this note, we propose weighted versions of the Lin-Wang (LW) test and investigate the performance of these weighted tests in a limited simulation study. Details are given in Section 2, and the simulation results are presented in Section 3. We give an example in Section 4 and conclude remarks in Section 5.

2. Methods

For consistency, we adhere to the notational conventions introduced by Lin and Wang [1]. We have survival data from two groups of subjects, the groups being labeled I and II, and are interested in comparing the survival distributions of the two groups. Events (failures or deaths) are observed at r distinct time points $t_1 < \cdots < t_r$ across the pooled groups. At time t_j, the number of observed failures in each of the two groups is denoted by d_{1j} for Group I and d_{2j} for Group II, and the numbers at risk just before time t_j are denoted by n_{1j} and n_{2j}, respectively, for $j = 1, 2, \ldots, r$. Consequently, at time t_j, there are $d_j = d_{1j} + d_{2j}$ failures out of $n_j = n_{1j} + n_{2j}$ subjects. Subjects may be censored during or at the end of the period of observation. A representative 2×2 contingency table of group by status at observed failure time t_j is given in Table 1.

We are interested in assessing the null hypothesis

H_0: the survival distributions of the two groups are identical versus the global alternative hypothesis.

H_1: the survival distributions of the two groups are not identical.

Lin and Wang introduced the quadratic statistic

$$\Delta = \sum_{j=1}^{r} \left[d_{1j} - E\left(d_{1j}\right) \right]^2 \tag{1}$$

for comparison of the two groups: they argued that Δ reflects the quadratic distance between the two underlying survival distributions hence should be sensitive to differences in either direction. They therefore based inference relating to H_0 on the standardized version of Δ, which they denoted as T^*.

Let us define a weighted version of Δ as

$$\Delta_w = \sum_{j=1}^{r} w_j * \left[d_{1j} - E(d_{1j}) \right]^2 \tag{2}$$

Table 1: Survival experience of the two groups at observed failure time t_j.

Group	Number of failures	Number of non-failures	Number at risk just before t_j
I	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
II	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Total	d_j	$n_j - d_j$	n_j

with arbitrary weights w_j, usually nonnegative. Our test statistic for assessing H_0 is the standardized version of Δ_W; namely,

$$T_w = \frac{\Delta_w - E(\Delta_w)}{\sqrt{\mathrm{Var}(\Delta_w)}}, \tag{3}$$

where $E(\Delta_W)$ and $\mathrm{Var}(\Delta_W)$ are calculated from the marginal hypergeometric distribution of the d_{1j}. In particular,

$$E(\Delta_w) = \sum_{j=1}^{r} w_j * \frac{n_{1j}n_{2j}d_j(n_j - d_j)}{n_j^2(n_j - 1)}, \tag{4}$$

and $\mathrm{Var}(\Delta_W)$ is given by

$$\begin{aligned}\mathrm{Var}(\Delta_w) = \sum_{j=1}^{r} w_j^2 * \Big\{ & E(d_{1j}^4) - 4E(d_{1j}^3)E(d_{1j}) \\ & + 6E(d_{1j}^2)\left[E(d_{1j})\right]^2 - 3\left[E(d_{1j})\right]^4 \\ & - \left[\mathrm{Var}(d_{1j})\right]^2 \Big\}.\end{aligned} \tag{5}$$

The raw moments of d_{ij} can be readily calculated from the following expression for the factorial moments:

$$E(d_{1j}^{(r)}) = \frac{n_{1j}^{(r)} d_j^{(r)}}{n_j^{(r)}}, \tag{6}$$

where $n^{(r)} = n^*(n-1) * \cdots * (n-r+1)$. For reference,

$$E(d_{1j}) = \frac{n_{1j}d_j}{n_j}, \tag{7}$$

$$\mathrm{Var}(d_{1j}) = \frac{n_{1j}n_{2j}d_j(n_j - d_j)}{n_j^2(n_j - 1)}, \tag{8}$$

$$E(d_{1j}^2) = \mathrm{Var}(d_{1j}) + \left[E(d_{1j})\right]^2, \tag{9}$$

$$E(d_{1j}^3) = 3E(d_{1j}^2) - 2E(d_{1j}) + \frac{n_{1j}^{(3)} d_j^{(3)}}{n_j^{(3)}}, \tag{10}$$

$$E(d_{1j}^4) = 6E(d_{1j}^3) - 11E(d_{1j}^2) + 6E(d_{1j}) + \frac{n_{1j}^{(4)} d_j^{(4)}}{n_j^{(4)}}. \tag{11}$$

We note in passing that there are typographical errors in the expressions for $E(d_{ij}^3)$ and $E(d_{ij}^4)$ in Lin and Wang [1].

Under the same assumptions as enumerated by Lin and Wang [1]; namely, the underlying failure times are independent, the censoring distributions (if any) for group I and group II are independent of each other, and of the respective survival distributions, the total number of observed failures and the distinct number of failure times are large, and the weights are positive and bounded; then T_w approximately follows a standard normal distribution. We are thus specifying the usual random censorship model, with further conditions to ensure approximate normality of T_w. For assessing the null hypothesis of equality of the underlying survival distributions of the two groups, Lin and Wang propose a two-sided test statistic based on T^*, and we will follow that convention with T_w.

3. Simulation Studies

In this section, we will investigate the empirical performance of weighted versions of the LW statistic, compared to the original (unweighted) LW statistic.

3.1. Empirical Type I Error. We first investigate achieved significance levels of the LW statistic and three weighted versions. Following LW, we generated two independent random samples from the exponential distribution with mean of 4. The censoring distribution is Uniform $(0, 20)$ in each group. The number of iterations in each simulation study is 5000. The empirical Type I error is calculated as the proportion of 5000 repeated random samples in which we reject the null hypothesis at the alpha = 0.05 significance level, under the assumption that T and weighted versions T_w have normal distributions, and two-sided tests are utilized. We report on three weighted versions of the LW statistic, delineated by different sets of weights w_j, $1 \le j \le r$: (i) $w_j = n_j$; (ii) $w_j = \sqrt{n_j}$; (iii) $w_j = 1/\mathrm{SD}(d_{1j})$, where $\mathrm{SD}(d_{1j}) = \sqrt{\mathrm{Var}(d_{1j})}$. The empirical Type I errors are given in Table 2.

In this limited simulation study the empirical Type I errors are quite close to the theoretical 0.05 value, for both the LW statistic and the weighted variants. The normal distribution seems an adequate approximation for the sample sizes investigated.

3.2. Empirical Power. Following LW, we undertook simulation studies comparing the empirical powers of the unweighted LW statistic with its weighted variants, under the three following scenarios.

Scenario 1. This scenario entails crossing survival curves. The LW specification is as follows. "In Group I the survival times follow an exponential distribution with mean of 6. In Group II the survival times follow an exponential distribution with mean of 2. However, if the survival time in Group II is greater than or equal to 1.5, then the survival time is regenerated to follow an exponential distribution with mean of 40. The censoring distribution is Uniform $(0, 20)$ in Group I and

Table 2: Empirical levels of the Lin-Wang test, and three weighted variants.

Sample sizes	LW	LW_{w_1}	LW_{w_2}	LW_{w_3}
(20, 20)	0.044	0.045	0.044	0.042
(30, 30)	0.048	0.053	0.053	0.048
(40, 40)	0.046	0.048	0.047	0.045
(50, 50)	0.051	0.053	0.050	0.049
(60, 60)	0.053	0.045	0.051	0.053
(70, 70)	0.049	0.049	0.051	0.047
(80, 80)	0.056	0.047	0.053	0.053
(90, 90)	0.051	0.049	0.050	0.050
(100, 100)	0.048	0.050	0.049	0.048

Notes: Sample sizes are given for group 1, followed by group 2. LW denotes the Lin-Wang test, and LW_{w_i} denotes the weighted version of the LW test, with weights w_i as described in the text. The underlying distributions of group 1 and group 2 were identical, as described in the text. The empirical levels of the two-sided test statistics were estimated from 5000 simulations, at nominal alpha level 0.05.

Uniform (0, 100) in Group II, which result in about 24% censoring rate in Group I and 18% in Group II, respectively."

Scenario 2. In this situation, the two survival curves are initially close, then cross, and diverge. The LW description is as follows. "In Group I the survival times follow an exponential distribution with mean of 4. In Group II the survival times follow an exponential distribution with mean of 3. However, if the survival time in Group II is greater than or equal to 4, then the survival time is regenerated to follow an exponential distribution with mean of 20. Also, censoring is assumed to occur randomly across the two groups. For each subject in the two groups, an independent Uniform (0, 1) random variable U is generated. In Group I, if U is less than 0.2, then the corresponding time point will be flagged as censored. Otherwise it is not censored. The censoring in Group II is created similarly but with a different rate. The censoring rate is 20% in Group I and 30% in Group II, respectively."

Scenario 3. Here, the proportional hazards assumption obtains. The LW specification is as follows. "The survival times follow an exponential distribution with means 2 and 5 in Groups I and II, respectively. The censoring mechanism is similar to that in Situation (Scenario 2), but this time with 20% censoring rate in Group I and 15% censoring rate in Group II, respectively."

The number of iterations in each simulation study is 5000. The estimated statistical power is calculated as the proportion of 5000 repeated random samples in which we reject the null hypothesis at the nominal alpha = 0.05 significance level, with two-sided test statistics. The weighted versions of the LW statistic are as above, namely, (i) $w_j = n_j$; (ii) $w_j = \sqrt{n_j}$; (iii) $w_j = 1/\mathrm{SD}(d_{1j})$, where $\mathrm{SD}(d_{1j}) = \sqrt{\mathrm{Var}(d_{1j})}$. Findings for the three scenarios are given in Tables 3, 4, and 5, respectively.

Interestingly, none of the procedures is dominant under every scenario. We might tend to favor the LW statistic under Scenario 1, the weighted version LW_{w3} under Scenario 2, and the weighted versions LW_{w1} and LW_{w2} under Scenario 3.

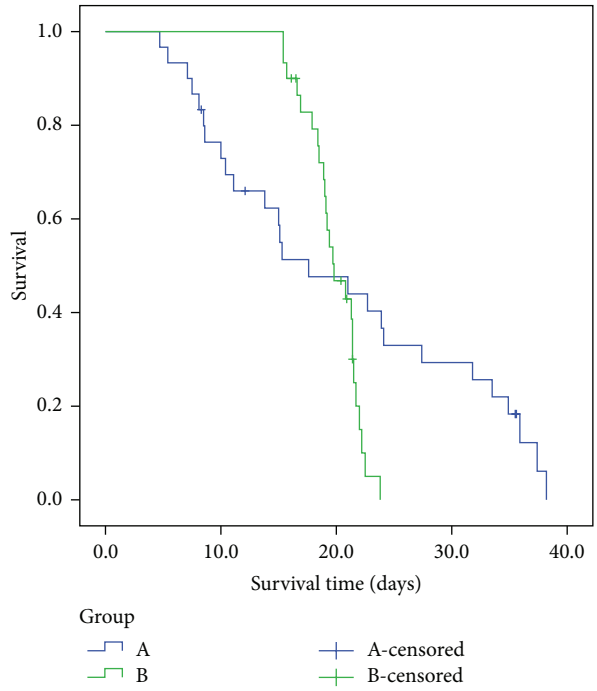


Figure 1: Kaplan-Meier survival curves for two groups of leukemic mice. Initial sample sizes were 30 per group. There were 4 censored observations in Group (A) and 5 in Group (B).

4. An Example

We will apply the various procedures to data arising from a cancer chemotherapy experiment, as explained in Koziol [2] and Koziol and Yuh [3]. Briefly, sixty leukemic mice were randomly subdivided into two groups of equal size; one group (Group (a)) was treated with a new investigative chemotherapeutic agent, and the other group (Group (b)) served as controls. Survival times of the two cohorts are given in Table 6, and Kaplan-Meier survival curves for the groups are depicted in Figure 1.

Clearly, we are in crossing hazards setting, and the logrank test and the generalized Wilcoxon test are not necessarily sensitive to this type of alternative. Indeed, with these data, the logrank chi-square statistic (with 1 d.f.) is 1.36 ($P = 0.24$), and the generalized Wilcoxon chi-square statistic is 1.12 ($P = 0.27$); we would fail to reject the hypothesis of equality of survival distributions for the two cohorts with either of these tests.

On the other hand, the LW statistic and its weighted variants all point to significantly different survival experiences in the two cohorts, with P values of 10^{-6} or smaller. In comparison, the omnibus Kolmogorov-Smirnov, Kuiper, and Cramér-von Mises statistics introduced by Koziol and Yuh [3] were also indicative of significantly different survival distributions but with more modest P values of 10^{-3}.

Table 3: Empirical powers of the Lin-Wang test, and three weighted variants, under Scenario 1.

Sample sizes	LW	LW_{w_1}	LW_{w_2}	LW_{w_3}
(20, 20)	0.406	0.269	0.331	0.407
(30, 30)	0.627	0.524	0.589	0.602
(40, 40)	0.813	0.772	0.81	0.774
(50, 50)	0.904	0.902	0.912	0.87
(60, 60)	0.952	0.968	0.965	0.922
(70, 70)	0.982	0.992	0.99	0.965
(80, 80)	0.99	0.998	0.996	0.981

Notes: Sample sizes are given for group 1, followed by group 2. LW denotes the Lin-Wang test, and LW_{w_i} denotes the weighted version of the LW test, with weights w_i as described in the text. The empirical powers of the two-sided test statistics were estimated from 5000 simulations, at nominal alpha level 0.05.

Table 4: Empirical powers of the Lin-Wang test, and three weighted variants, under Scenario 2.

Sample sizes	LW	LW_{w_1}	LW_{w_2}	LW_{w_3}
(20, 20)	0.089	0.039	0.053	0.097
(30, 30)	0.157	0.046	0.084	0.167
(40, 40)	0.222	0.057	0.116	0.239
(50, 50)	0.314	0.077	0.169	0.339
(60, 60)	0.402	0.106	0.225	0.433
(70, 70)	0.484	0.122	0.273	0.511
(80, 80)	0.549	0.151	0.338	0.583
(90, 90)	0.612	0.173	0.379	0.644
(100, 100)	0.675	0.212	0.431	0.700

Notes: Sample sizes are given for group 1, followed by group 2. LW denotes the Lin-Wang test, and LW_{w_i} denotes the weighted version of the LW test, with weights w_i as described in the text. The empirical powers of the two-sided test statistics were estimated from 5000 simulations, at nominal alpha level 0.05.

Table 5: Empirical powers of the Lin-Wang test, and three weighted variants, under Scenario 3.

Sample sizes	LW	LW_{w_1}	LW_{w_2}	LW_{w_3}
(20, 20)	0.433	0.430	0.435	0.387
(30, 30)	0.600	0.630	0.628	0.532
(40, 40)	0.728	0.775	0.767	0.647
(50, 50)	0.822	0.878	0.872	0.753
(60, 60)	0.889	0.929	0.925	0.827
(70, 70)	0.931	0.970	0.962	0.884
(80, 80)	0.964	0.985	0.982	0.933

Notes: Sample sizes are given for group 1, followed by group 2. LW denotes the Lin-Wang test, and LW_{w_i} denotes the weighted version of the LW test, with weights w_i as described in the text. The empirical powers of the two-sided test statistics were estimated from 5000 simulations, at nominal alpha level 0.05.

5. Concluding Remarks

The logrank test as described in Section 2 should be ascribed to Mantel [4]: Mantel brilliantly intuited that the Mantel-Haenszel (MH) statistic [5] for assessing association across independent 2×2 tables could be applied to survival data, by constructing a 2×2 table as in Table 1 at each event (death) time then combining the resulting 2×2 tables as in the MH procedure.

Correspondingly, our incorporation of weights into the LW statistic as described in Section 2 is not new: our motivation devolves from similar introduction of weights into the Mantel formulation of the logrank statistic, by Tarone and Ware [6] and Leurgans [7] among others. And, anticipating the findings in Section 3, these investigators have shown that the weights can enjoy improved power properties over the unweighted MH statistics in various settings. We remark that calculation of the LW statistic is rather computationally intensive; but incorporation of weights should cause no additional computational difficulties. Optimal choice of weights remains an open issue, which we are currently pursuing.

The generalized Wilcoxon test and the logrank test are perhaps the best known and most commonly used procedures for the comparison of two survival distributions

Table 6: The clinical data for sixty leukemic mice which were randomly subdivided into two groups (Group A and Group B) of equal size. "1" indicates the censored data.

Group A		Group B	
Survival (days)	Censoring	Survival (days)	Censoring
4.7	0	15.4	0
5.4	0	15.4	0
7.1	0	15.7	0
7.5	0	16.1	1
8.1	0	16.5	1
8.3	1	16.6	0
8.5	0	16.9	0
8.6	0	17.9	0
10	0	18.4	0
10.4	0	18.5	0
11.1	0	18.9	0
12.1	1	19	0
13.8	0	19.1	0
15	0	19.2	0
15.1	0	19.4	0
15.3	0	19.7	0
17.6	0	19.8	0
21	0	20.4	1
22.7	0	20.8	0
23.9	0	20.9	1
24.1	0	21.3	0
27.4	0	21.4	0
31.8	0	21.4	0
33.5	0	21.4	1
34.9	0	21.5	0
35.5	1	21.7	0
35.6	1	22	0
35.9	0	22.2	0
37.4	0	22.5	0
38.2	0	23.8	0

with observations subject to random censorship. Mantel [4] and others recognized, however, that these tests may not be appropriate whenever the alternative of interest is not that the one survival distribution is stochastically larger than the other but merely that the distributions are not equal. Crossing hazards are an example of nonstochastic ordering of survival distributions. For testing equality against such alternatives, Koziol [2] proposed a two-sample Cramér-von Mises type statistic based on the product-limit estimates of the individual survival distributions, and later Koziol and Yuh [3] introduced Kolmogorov-Smirnov and Kuiper as well as Cramér-von Mises statistics for the same omnibus two-sample testing problem. The LW statistic is more closely attuned to the logrank test than these omnibus procedures; and, as seen in the example, the LW statistics may be more sensitive to crossing hazards alternatives.

It should be noted that Mantel [4] also proposed a modification of the Mantel logrank test, appropriate for crossing hazards: Mantel suggested that one construct a "chi-squared" statistic at each event time as in Table 1, sum these individual statistics over the event times, and then treat the resulting sum as an approximate chi-square random variable with n degrees of freedom, n being the number of tables (distinct event times). We explored this statistic in simulation studies, but regrettably we cannot recommend this statistic, due to decreased power relative to the other statistics reported herein, and the tenuous assumption that a chi-square distribution for this statistic is adequate (though with larger sample sizes, a normal approximation might be invoked).

Conflict of Interests

The authors declare that they have no competing interests.

Authors' Contributions

James A. Koziol conceived of the study. Zhenyu Jia and James A. Koziol carried out the data analysis and drafted the paper. Both authors read and approved the final paper.

Acknowledgments

This work was supported by the National Cancer Institute Early Detection Research Network (EDRN) Consortium Grant no. U01 CA152738. Many years ago, James A. Koziol was mentored by Nathan Mantel at the National Cancer Institute and will be forever indebted to him.

References

[1] X. Lin and H. Wang, "A new testing approach for comparing the overall homogeneity of survival curves," *Biometrical Journal*, vol. 46, no. 5, pp. 489–496, 2004.

[2] J. A. Koziol, "A two sample Cramer-von Mises test for randomly censored data," *Biometrical Journal*, vol. 20, no. 6, pp. 603–608, 1978.

[3] J. A. Koziol and Y. S. Yuh, "Omnibus two-sample test procedures with randomly censored data," *Biometrical Journal*, vol. 24, no. 8, pp. 743–750, 1982.

[4] N. Mantel, "Evaluation of survival data and two new rank order statistics arising in its consideration," *Cancer Chemotherapy Reports*, vol. 50, no. 3, pp. 163–170, 1966.

[5] N. Mantel and W. Haenszel, "Statistical aspects of the analysis of data from retrospective studies of disease," *Journal of the National Cancer Institute*, vol. 22, no. 4, pp. 719–748, 1959.

[6] R. E. Tarone and J. Ware, "On distribution free tests for equality of survival distributions," *Biometrika*, vol. 64, no. 1, pp. 156–160, 1977.

[7] S. Leurgans, "Three classes of censored data rank tests: strengths and weaknesses under censoring," *Biometrika*, vol. 70, no. 3, pp. 651–658, 1983.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 316935, 9 pages
http://dx.doi.org/10.1155/2014/316935

Research Article

Logic Regression for Provider Effects on Kidney Cancer Treatment Delivery

Mousumi Banerjee,[1,2] Christopher Filson,[3] Rong Xia,[1] and David C. Miller[2,3]

[1] *Department of Biostatistics, School of Public Health, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109, USA*
[2] *Center for Healthcare Outcomes & Policy, University of Michigan, 2800 Plymouth Road, Ann Arbor, MI 48105, USA*
[3] *Department of Urology, University of Michigan, 1500 E Medical Center Drive, Ann Arbor, MI 48109, USA*

Correspondence should be addressed to Mousumi Banerjee; mousumib@umich.edu

Received 9 January 2014; Accepted 28 February 2014; Published 27 March 2014

Academic Editor: Zhenyu Jia

In the delivery of medical and surgical care, often times complex interactions between patient, physician, and hospital factors influence practice patterns. This paper presents a novel application of logic regression in the context of kidney cancer treatment delivery. Using linked data from the National Cancer Institute's (NCI) Surveillance, Epidemiology, and End Results (SEER) program and Medicare we identified patients diagnosed with kidney cancer from 1995 to 2005. The primary endpoints in the study were use of innovative treatment modalities, namely, partial nephrectomy and laparoscopy. Logic regression allowed us to uncover the interplay between patient, provider, and practice environment variables, which would not be possible using standard regression approaches. We found that surgeons who graduated in or prior to 1980 despite having some academic affiliation, low volume surgeons in a non-NCI hospital, or surgeons in rural environment were significantly less likely to use laparoscopy. Surgeons with major academic affiliation and practising in HMO, hospital, or medical school based setting were significantly more likely to use partial nephrectomy. Results from our study can show efforts towards dismantling the barriers to adoption of innovative treatment modalities, ultimately improving the quality of care provided to patients with kidney cancer.

1. Introduction

Open radical nephrectomy has long been the standard treatment for patients with early-stage kidney cancer [1]. In recent years, however, easier convalescence and equivalent cancer control established laparoscopy as an alternative standard of care for most patients treated with radical nephrectomy [1–3]. Studies have also demonstrated that, for patients with small renal masses, partial instead of radical nephrectomy achieves identical cancer control while better preserving long-term renal function and reducing overtreatment of benign or clinically indolent tumors [4–7]. However, despite their potential advantages, the adoption of laparoscopy and partial nephrectomy have been relatively slow and asymmetric in the population [3, 8].

Earlier studies have shown that individual surgeon characteristics and their practice environments largely influence the use of laparoscopy and partial nephrectomy [9]. These studies are based on logistic regression models, a member of the generalized linear model family suitable for data with a binary outcome (e.g., use versus nonuse of laparoscopy). Logistic regression focuses on identification of *main effects*. While interactions can be assessed using logistic regression, these interactions need to be known a priori and specified as input variables in the model. *Discovery of interactions* is therefore difficult using logistic regression. We hypothesize that surgeon characteristics may not have uniform effect on the adoption of laparoscopy and partial nephrectomy across practice environments. For example, use of advanced techniques may vary among recently trained surgeons depending on the surgeon's affiliation with an academic hospital or NCI-designated cancer center, suggesting a potential interaction between year of medical school graduation and practice setting.

Logic regression is an adaptive classification and regression procedure [10], initially developed to uncover and

measure the importance of interacting factors in genetic association studies [11, 12]. There are many approaches based on classification methods such as CART and Random Forests [13–15] that allow measuring the importance of a single predictor. But none of these methods can directly quantify the importance of combinations of several predictors. Logic regression uses the predictors as inputs into the model while still enabling one to identify combinations of predictors and quantify the importance of these interactions.

In general, logic regression can be used in any setting, when the interaction between the predictors is of primary interest. Logic regression searches for Boolean (logical) combinations of the original predictors that best explain the variability in the outcome variable and, thus, reveals variables and interactions that are associated with the response and/or have predictive capabilities. Given a set of binary predictors, one creates new predictors such as "X_1, X_2, X_3, and X_4 are true" or "X_5 or X_6 but not X_7 is true." In more specific terms, the goal is to try to fit regression models of the form $\text{logit}[P(Y = 1)] = b_0 + b_1 L_1 + \cdots + b_p L_p$, where $P(Y = 1)$ is the probability that the binary outcome is 1, and L_j is any Boolean expression of the predictors. The L_j and b_j are estimated simultaneously using a stochastic optimization algorithm [16].

The goal of this paper is to introduce logic regression as a novel method for discovering interactions, specifically, Boolean combinations of factors that potentially discriminate users of partial nephrectomy or laparoscopy from nonusers. Characterizing providers who are actually using (or not using) these techniques is needed to show education- and/or policy-based interventions designed to increase utilization of these advanced surgical techniques. Given that logic expressions are embedded in a generalized linear regression framework and therefore naturally adaptable to other outcome types (e.g., numeric and time-to-event data); the method has broad scope of application in health services and outcomes research.

2. Materials and Methods

2.1. Data Source. We used data from the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program and the Centers for Medicare and Medicaid Services (Medicare) to identify patients diagnosed with incident kidney cancer from 1995 to 2005. SEER is a population-based cancer registry that collects data regarding incidence, treatment, and mortality. The demographic composition, cancer incidence, and mortality trends in the SEER registries are representative of the entire United States population. The Medicare Program provides primary health insurance for 97% of the United States population aged 65 years and older, and linkage to Medicare claims is achieved for >90% of SEER cases over age 65 [17].

2.2. Cohort Identification and Assignment of Surgical Procedure. We identified 15,744 patients diagnosed with nonurothelial, nonmetastatic kidney cancer from 1995 to 2005. For this group of patients, we searched inpatient and physician claims to identify kidney cancer-specific diagnosis and procedure codes (list of codes available from authors upon request). We excluded patients who lacked claims denoting surgical treatment for kidney cancer, patients with multiple hospitalizations for direct open or partial nephrectomy, patients whose claims suggested the presence of bilateral tumors at diagnosis, and patients operated by a nonurologic specialty physician. This process yielded a cohort of 11,918 cases. We applied a validated claim-based algorithm to assign each patient to one of four mutually exclusive surgical categories: (1) open radical nephrectomy (ORN) ($n = 8029$), (2) open partial nephrectomy (OPN) ($n = 1380$), (3) laparoscopic radical nephrectomy (LRN) ($n = 2082$), and (4) laparoscopic partial nephrectomy (LPN) ($n = 427$).

As validation, we assessed the level of concordance between our claim-based algorithm and the type of cancer-directed surgery specified for each patient in the SEER data file (Patient Entitlement and Diagnosis Summary File). Although SEER does not collect data regarding whether the surgical approach was open or laparoscopic, we observed 97% agreement for the assignment of partial versus radical nephrectomy ($\kappa = 0.83$). Also, we identified relevant surgical pathology claims within 30 days of the index admission for more than 95% of analyzed cases, thus supporting the occurrence of cancer-directed surgery. As a final step, we externally validated our algorithm by comparing procedure assignments based on Medicare claims with the surgery specified in actual operative reports of 549 cases from the Los Angeles Cancer Surveillance Program. Overall, the claims-based algorithm assigned the correct surgical procedure (ORN, OPN, LRN, or LPN) for 97% of patients in the validation sample ($\kappa = 0.91$). We observed equally high concordance for identification of laparoscopic versus open surgery ($\kappa = 0.87$) and for classification of partial versus radical nephrectomy ($\kappa = 0.93$).

2.3. Patient-Level Covariates. For each patient in the analytic cohort, we used SEER data to determine demographic and cancer-specific information (i.e., age at surgery, gender, race/ethnicity, marital status, tumor size, tumor grade, histology, and laterality). Based on patient-level zip codes, we assigned patients to one of three socioeconomic strata [18]. We measured preexisting comorbidity by using a modified Charlson Index based on claims submitted during the 12 months prior to the kidney cancer surgery [19, 20].

2.4. Primary Surgeon and Surgeon-Level Covariates. To identify the primary surgeon for each case, we used encrypted Unique Physician Identifier Numbers (UPIN) submitted with Medicare physician claims. We linked the comprehensive list of surgeon UPINs to the American Medical Association (AMA) Physician Masterfile, which contains demographic, educational, and certification information for over one million residents and physicians in the United States. Using AMA data, we determined surgeon age, gender, year of medical school graduation, and practice size. We assigned each surgeon a rural-urban commuting area (RUCA) code

based on an established classification scheme using the zip code of the primary office address [21]. We determined academic affiliation (major, minor, or no academic affiliation) based on the methods described by Shahinian et al. [22]. We also determined each surgeon's average annual nephrectomy (partial or radical) volume using claims from 1995 to 2005. We empirically defined high-volume surgeons as those performing at least 3 annual cancer-related nephrectomies among the SEER-Medicare population (83rd percentile). This measure of case volume may not reflect the total number of nephrectomies performed by a provider: it fails to account for surgeries among younger (non-Medicare-eligible) patients, Medicare HMO enrollees, and/or fee-for-service Medicare participants who reside outside of the SEER registries. Finally, we determined each surgeon's association with a National Cancer Institute- (NCI-) designated cancer center based on whether or not they performed at least one radical nephrectomy at a hospital carrying this designation.

2.5. Statistical Methods. Before fitting logic regression models, we performed several univariate analyses. We used Chi-square tests to evaluate the level of association between surgical procedure and various patient-level covariates and to assess the statistical significance of temporal surgical trends. For the subsequent modeling, we defined two binary endpoints as follows: (1) use of partial nephrectomy (i.e., OPN+LPN versus ORN+LRN) and (2) use of laparoscopy among patients who underwent radical nephrectomy (i.e., LRN versus ORN).

The classification algorithm used in this study is logic regression, an adaptive regression methodology developed by Ruczinski et al. [10]. In the logic regression framework, given a set of binary covariates X, the goal is to create new, better predictors for the response by constructing Boolean combinations of the binary covariates. For example, if the response is binary, the goal is to find decision rules such as "if X_1, X_2, X_3, and X_4 are true," or "X_5 or X_6 but not X_7 is true," then the response is more likely to be in class 0. Boolean combinations of the covariates, called logic trees, are represented graphically as a set of and-or rules. Logic regression searches for Boolean combinations of predictors in the entire space of such combinations, while being completely embedded in a regression framework, where the quality of the models is determined by an appropriate score function for the regression class.

Let $X_1, X_2, \ldots, X_k$ be binary (0/1) predictors and let Y be the response. In our setting, X's correspond to patient, physician, and practice environment variables, and Y represents binary outcomes (use of partial nephrectomy: yes/no; use of laparoscopy: yes/no) each of which is modeled separately using binomial deviance as the score function. For a given set of Boolean expressions, an example of which was given in Section 1, the logic regression model is a logistic regression model with those Boolean expressions as covariates. Specifically, we denote a Boolean expression with the binary variable L, where $L = 1$ is true and $L = 0$ is false. The model is written as

$$\text{logit}P\left(Y = 1 \mid L_1, \ldots, L_p\right) = \beta_0 + \beta_1 L_1 + \cdots + \beta_p L_p, \quad (1)$$

where L_j is a Boolean combination of the predictors X_i's. The goal is to find Boolean expressions in (1) that minimize the binomial deviance, estimating the parameters β_j simultaneously with the search for the Boolean expressions L_j. This is what distinguishes logic regression from simple logistic regression with binary covariates, that is, that the fitting algorithm both defines "covariates" for model (1) (using predictor data) and estimates the regression coefficients simultaneously. The output from logic regression is represented as a series of trees, one for each Boolean predictor, L_j, and the associated regression coefficient.

CART is another tree-based method for modeling binary data [15]. The classification rule is displayed as a tree whose leaves are the two classes of interest (e.g., use versus nonuse of partial nephrectomy or laparoscopy) and whose branches correspond to dichotomized covariates. Each leaf is reached by one or more paths through the tree; to reach the leaf, all conditions along the path must be satisfied. Thus, a classification tree can be thought of as the collection of all paths that reach a leaf predicting use of treatment. Therefore, any classification tree can be written as a Boolean combination of covariates, as can a logic regression tree. However, there are some Boolean expressions which can be very simply represented as logic trees, but which require fairly complicated classification trees [10]. It is this simplicity of logic trees which we hope to exploit in order to produce easily interpretable characterizations of individuals who have a high likelihood of using the specific surgical treatment.

In logic regression, the challenge is to find good candidates for the logic term L_j, as the collection of all Boolean expressions is enormous. Using a tree-like representation for logic expressions, we adaptively select this term using a simulated annealing algorithm [16]. In our setting leaves of each tree are the threshold conditions for each covariate, and the root and knots of the tree are the Boolean (and-or) operators. Simulated annealing is a stochastic optimization algorithm. At each step a possible operation on the current tree, such as adding or removing a knot, is proposed at random. This operation is always accepted if the new logic tree has a better score than the old logic tree; otherwise, it is accepted with a probability that depends on the difference between the scores of the old and the new tree and the stage of the algorithm. Properties of the simulated annealing algorithm depend heavily on Markov chain theory and thus on the set of operations that can be applied to logic trees.

The complexity of a logic regression model is defined by the number of logic trees (p in (1)) and the number of variables, or leaves, that make up a tree. As with any adaptive regression methodology, larger models (those with more trees and leaves) typically fit better than smaller models. To avoid overfitting, in this paper we chose the model size using a cross-validation approach. We varied model complexity from 1 to 4 trees (corresponding to the p in (1)) and the number of leaves that make up a tree from 1 to 15. We randomly divided our data into ten subsets, such that each subset consisted of one-tenth of the "treatment" and the "control" (RN for the first endpoint and ORN for the second endpoint) groups. Of the ten subsets, we used nine subsets as training data and the remaining single subset as validation data for testing the

model. We used the training data to develop the logic models using simulated annealing algorithm and then estimated the deviance based on the test data. This process was repeated ten times, with each of the ten subsets used exactly once as validation data (tenfold cross validation). The results from the ten folds were then averaged to produce a single deviance score for each model. To reduce variability, this procedure (splitting the data into ten parts, developing logic rules on the training data, and estimating the deviance based on the test data) was repeated 15 times, with different random splits of the whole dataset for each run. The deviance scores were averaged over the 15 rounds of cross-validation, and the model with the smallest average deviance was selected. Results presented correspond to the run yielding value for the test set based model deviance that was closest to its average across the 15 runs. This run was selected so as to provide results for what might be considered a typical rather than an extreme split of the data into test and training sets.

Logic regression requires binary predictor variables, so we recoded variables into binary forms. Categorical covariates were coded as a set of indicator variables for each level of the covariate. For example, marital status was analyzed as married versus others, and race was coded as a set of indicator variables based on the categories Caucasian, African American, Hispanic, and others. Continuous and ordinal covariates were coded as a series of threshold indicators based on a priori knowledge about the variables. For example, tumor size was categorized into two clinically relevant groups based on a 4 cm threshold; patient's age at surgery and Charlson comorbidity index were each coded as a series of threshold indicators based on five-year age intervals 65–69, 70–74, 75–79, 80–84, and ≥85 years and 0, 1, and ≥2 comorbid conditions, respectively. Each of the surgeon variables, that is, surgeon's year of medical school graduation, age, practice structure, and academic affiliation, was also coded as a series of threshold indicators based on the categories prior to 1960, 1961–1970, 1971–1980, 1981–1990, and 1991 and after; <40, 40–49, 50–59, and ≥60 years; solo or two person practice, group practice, HMO or hospital based practice, medical school, and others; and none, minor, and major affiliation, respectively.

3. Results

We identified a total of 11,918 Medicare beneficiaries who underwent surgery for an incident kidney cancer diagnosed between 1995 and 2005. Table 1 presents demographic and clinical characteristics of patients in the analytic sample. During the study interval, 1807 patients (15.2%) underwent partial nephrectomy (427 performed laparoscopically), and 10,111 patients (84.8%) underwent radical nephrectomy (2082 performed laparoscopically). We observed differences in treatment patterns according to patient age, gender, race/ethnicity, marital status, socioeconomic status, tumor size, tumor grade, and histology (Table 1).

From 1995 to 2005, the annual proportion of patients who underwent partial nephrectomy increased from 8.5% to 21.3% ($P < 0.0001$); for patients who had tumors that measured ≤4 cm, the proportion rose from 14.4% to 37.1% ($P < 0.0001$). Among patients treated with radical nephrectomy, the annual proportion of laparoscopy use increased from 1.3% in 1995 to 44.1% in 2005 ($P < 0.0001$). For patients whose tumors measured ≤4 cm, the annual proportion of laparoscopy use increased from 1.6% to 52.9%; for patients with larger tumors, this proportion increased from 1.2% to 39.3% (P values < 0.0001).

We identified 2088 primary surgeons who performed 11,918 kidney cancer surgeries during the study interval (median 4 cases). Of these, 2019 surgeons performed 10,111 radical nephrectomies (median 3 cases; range 1–84). During the same interval, 842 surgeons performed 1,807 partial nephrectomies (median 1 case; range 1–29). Of the 2019 surgeons who performed the radical nephrectomies, 720 operated laparoscopically on 2,082 patients (median 2 cases; range 1–55). We observed differences in treatment patterns according to provider age, gender, year of medical school graduation, annual nephrectomy volume, practice size, rural/urban status, academic affiliation, and NCI cancer center designation (Table 2).

Figure 1 displays results of the logic regression to determine optimal combination rules for use of partial nephrectomy based on a two-tree model. The first tree, L_1, is entirely described by tumor size. The estimated odds ratio associated with this tree is 5.9 (95% CI 4.7–7.4), suggesting that tumor size ≤4 cm is associated with almost six times higher odds of partial nephrectomy. This finding is concurrent with previous reports in the literature documenting tumor size as a strong predictor of partial nephrectomy [8, 9]. Interestingly, the second tree, L_2, involves practice environment characteristics. This tree (L_2) indicates that not having major academic affiliation, or not practising in HMO, hospital, medical school based setting is associated with lower odds ratio of partial nephrectomy. The estimated odds ratio associated with L_2 is 0.30 (95% CI 0.23–0.39), suggesting that, as a group, those satisfying L_2 are estimated to have a 70% lower odds of using partial nephrectomy compared to those who do not satisfy the tree. In other words, patients treated by surgeons who have major academic affiliation and are in HMO, hospital, or medical school based practice setting are 3.3 times more likely to undergo partial nephrectomy than their counterparts.

Figure 2 displays results of the logic regression to determine optimal combination rules for use of laparoscopic radical nephrectomy based on a three-tree model. The first tree, L_1, is entirely described by academic affiliation. The estimated odds ratio associated with this tree is 2.12 (95% CI 1.71–2.63), suggesting that surgeon's affiliation with a major academic center is associated with two times higher odds of a laparoscopic radical nephrectomy. The second tree, L_2, involves a combination of patient and surgeon characteristics. This tree (L_2) indicates that having larger tumors (>4 cm) or having a surgeon who graduated in or prior to 1980 or practising in nongroup settings (solo or two person) is associated with a lower odds of laparoscopic procedure. The estimated odds ratio associated with L_2 is 0.38 (95% CI 0.29–0.48), suggesting that, as a group, those satisfying L_2 are estimated to have a 62% lower odds of laparoscopic radical nephrectomy compared to those who do not satisfy the tree.

Table 1: Distribution of patient and tumor characteristics by surgical procedures (1995–2005).

	Total	LPN	LRN	OPN	ORN	
	n	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*P* value
	11,918	427 (3.6)	2082 (17.5)	1380 (11.6)	8029 (67.3)	
Age at surgery (years)						0.0001
65–69	3127	131 (4.2)	530 (16.9)	431 (13.8)	2035 (65.1)	
70–74	3423	122 (3.6)	536 (15.7)	426 (12.5)	2339 (68.3)	
75–79	3024	98 (3.2)	579 (19.2)	354 (11.7)	1993 (65.9)	
80–84	1721	59 (3.4)	300 (17.4)	139 (8.1)	1223 (71.1)	
≥85	623	17 (2.7)	137 (22.0)	30 (4.8)	439 (70.5)	
Race/ethnicity						0.0001
Caucasian	9884	344 (3.5)	1757 (17.8)	1158 (11.7)	6625 (67.0)	
African-American	878	33 (3.8)	154 (17.5)	103 (11.7)	588 (67.0)	
Hispanic	719	27 (3.8)	81 (11.3)	70 (9.7)	541 (75.2)	
Other or Unknown	437	23 (5.3)	90 (20.6)	49 (11.2)	275 (62.9)	
Gender						0.0001
Male	6882	274 (3.9)	1134 (16.5)	850 (12.4)	4624 (67.2)	
Female	5036	153 (3.0)	948 (18.8)	530 (10.5)	3405 (67.6)	
Marital status						0.005
Yes	7499	294 (3.9)	1274 (17.0)	901 (12.0)	5030 (67.1)	
No	4419	133 (3.0)	808 (18.3)	479 (10.8)	2999 (67.9)	
Socioeconomic status						0.0001
Low	3808	134 (3.5)	603 (15.8)	424 (11.1)	2647 (69.5)	
Intermediate	3899	135 (3.5)	633 (16.2)	386 (9.9)	2745 (70.4)	
High	4196	158 (3.8)	846 (20.2)	568 (13.5)	2624 (62.5)	
Charlson comorbidity score						0.38
0	6842	241 (3.5)	1186 (17.3)	794 (11.6)	4621 (67.5)	
1	2847	104 (3.7)	512 (18.0)	313 (11.0)	1918 (67.4)	
≥2	1904	74 (3.9)	345 (18.1)	246 (12.9)	1239 (65.1)	
Tumor size (cm)						0.0001
≤4	5188	352 (6.8)	949 (18.3)	1035 (20.0)	2852 (54.9)	
>4	6401	51 (0.8)	1101 (17.2)	286 (4.5)	4963 (77.5)	
Tumor histology						0.0001
Clear cell	10000	301 (3.0)	1682 (16.8)	1042 (10.4)	6975 (69.8)	
Papillary	888	77 (8.7)	200 (22.5)	170 (19.1)	441 (49.7)	
Chromophobe	391	24 (6.1)	107 (27.4)	82 (21.0)	178 (45.5)	
Other	639	25 (3.9)	93 (14.6)	86 (13.5)	435 (68.1)	

The third tree, L_3, is characterized by a combination of surgeon and practice environment variables. This tree (L_3) indicates that low volume surgeons in a non-NCI hospital, surgeons in rural environment, or surgeons who graduated in or prior to 1980 despite having some academic affiliation have a lower odds for laparoscopic procedure (odds ratio = 0.29, 95% CI 0.23–0.38).

We also performed CART analyses of our data (results not shown) for both the partial and laparoscopic radical nephrectomy endpoints. For partial nephrectomy, the CART tree yielded subgroups characterized by tumor size and surgeon's academic affiliation. As observed before, patients with tumor size > 4 cm were less likely to undergo partial nephrectomy compared to those with smaller tumors. For the latter group (tumor size ≤4 cm), surgeons with major academic affiliation had higher propensity for partial nephrectomy compared to those with minor or no academic affiliation. The area under the ROC curve for the CART tree was 0.72, compared to 0.77 for the logic regression model. For laparoscopic radical nephrectomy, the CART tree yielded subgroups characterized by surgeon's academic affiliation, year of medical school graduation, and annual surgeon volume. High volume surgeons who graduated after 1980 and were affiliated with a major academic center had the highest propensity towards laparoscopic procedure. Surgeons with minor or no academic affiliation had the lowest propensity towards laparoscopic procedure. Interestingly, despite having major academic affiliation surgeons who graduated in or prior to 1980 had only a slightly higher propensity towards laparoscopic procedure compared to surgeons with minor or

Table 2: Distribution of surgeon and practice environment characteristics by surgical procedures (1995–2005).

	Total	LPN	LRN	OPN	ORN	
	n	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*P* value
	11,918	427 (3.6)	2082 (17.5)	1380 (11.6)	8029 (67.3)	
Surgeon age (years)						0.0001
<40	2553	147 (5.8)	774 (30.3)	271 (10.6)	1361 (53.3)	
40–49	4034	170 (4.2)	728 (18.1)	440 (10.9)	2696 (66.8)	
50–59	3710	85 (2.3)	427 (11.5)	458 (12.4)	2740 (73.9)	
≥60	1621	25 (1.5)	153 (9.4)	211 (13.0)	1232 (76.0)	
Surgeon gender						0.0001
Male	11684	419 (3.6)	2036 (17.4)	1364 (11.7)	7865 (67.3)	
Female	234	8 (3.4)	46 (19.7)	16 (6.8)	164 (70.1)	
Annual nephrectomy volume						0.0001
Bottom 25%	2279	34 (1.5)	209 (9.2)	230 (10.1)	1806 (79.3)	
2nd 25%	3474	77 (2.2)	458 (13.2)	370 (10.7)	2569 (73.9)	
3rd 25%	3141	107 (3.4)	519 (16.5)	320 (10.2)	2195 (69.9)	
Top 25%	3024	209 (6.9)	896 (29.6)	460 (15.2)	1459 (48.3)	
Year of medical school graduation						0.0001
<1960	346	2 (0.6)	9 (2.6)	48 (13.9)	287 (82.9)	
1961–1970	2488	25 (1.0)	176 (7.1)	301 (12.1)	1986 (79.8)	
1971–1980	3568	89 (2.5)	437 (12.3)	403 (11.3)	2639 (73.9)	
1981–1990	3705	166 (4.5)	738 (19.9)	431 (11.6)	2370 (63.9)	
>1991	1811	145 (8.0)	722 (39.9)	197 (10.9)	747 (41.3)	
Practice size						0.0001
Solo or two-person	3200	36 (1.1)	284 (8.9)	279 (8.7)	2601 (81.3)	
Group practice	6619	274 (4.1)	1368 (20.7)	709 (10.7)	4268 (64.5)	
HMO or hospital-based	631	29 (4.6)	100 (15.9)	141 (22.4)	361 (57.2)	
Medical school	484	34 (7.0)	98 (20.3)	125 (25.8)	227 (46.9)	
Other/unclassified	984	54 (5.5)	232 (23.6)	126 (12.8)	572 (58.1)	
Academic affiliation						0.0001
None	4195	88 (2.1)	660 (15.7)	385 (9.2)	3062 (72.9)	
Minor	4408	127 (2.9)	740 (16.8)	420 (9.5)	3121 (70.8)	
Major	3201	207 (6.5)	668 (20.9)	561 (17.5)	1765 (55.1)	
Rural/urban status						0.0001
Urban	11093	412 (3.7)	1992 (17.9)	1318 (11.9)	7371 (66.5)	
Rural	823	15 (1.8)	89 (10.8)	61 (7.4)	658 (79.9)	
Cancer Center Affiliation						0.0001
No	10793	322 (2.9)	1861 (17.2)	1125 (10.4)	7485 (69.4)	
Yes	1096	102 (9.3)	219 (19.9)	252 (22.9)	523 (47.7)	

no academic affiliation. The area under the ROC curve for the CART tree was 0.61, compared to 0.71 for the logic regression model.

4. Conclusions

The principal finding from this study was our ability to uncover the interplay between patient, provider, and practice environment variables towards adoption of partial nephrectomy and laparoscopy. Through the use of logic regression we were able to uncover interactions that would not have been detected by standard logistic regression approach. Our findings demonstrate that the adoption of laparoscopic radical nephrectomy is particularly influenced by complex combinations of surgeon and practice environment characteristics, rather than simple "main effects." More specifically, our results suggest that patients treated by surgeons who graduated in or prior to 1980 despite having some academic affiliation, low volume surgeons in a non-NCI hospital, or surgeons in rural environment were significantly less likely to use laparoscopic radical nephrectomy. Although less dramatic, the adoption of partial nephrectomy is also influenced by combination of tumor and practice environment characteristics. Collectively, these findings highlight the rich contextual interactions that influence urologist's adoption of

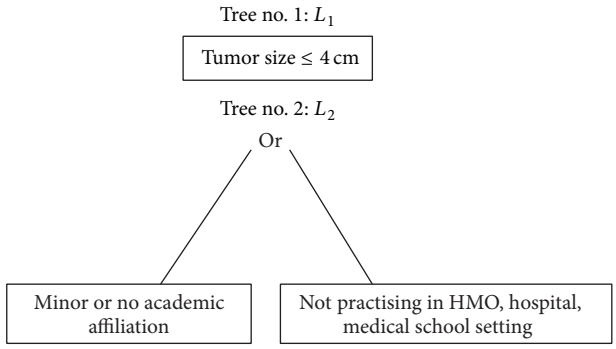


Figure 1: Two-tree model for use of partial nephrectomy. The odds ratio associated with L_1 is 5.9 (95% CI 4.7–7.4) and that with L_2 is 0.30 (95% CI 0.23–0.39).

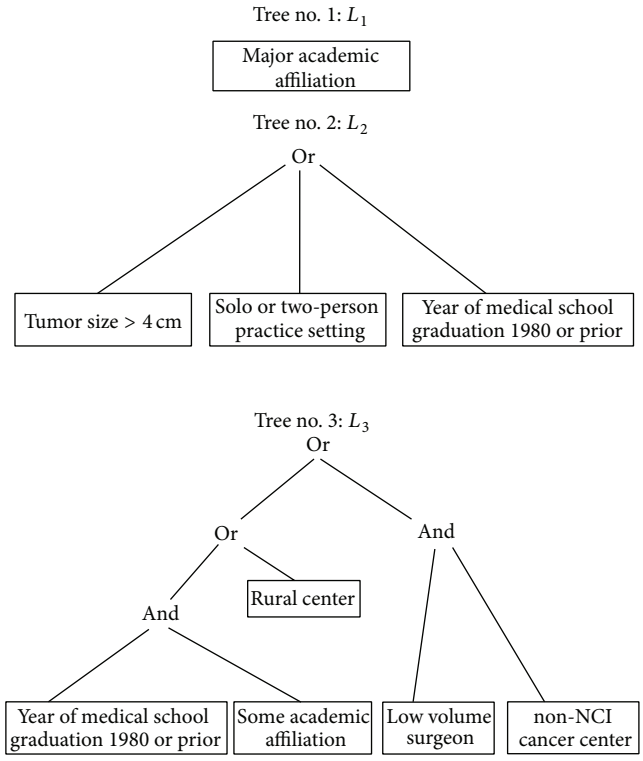


Figure 2: Three-tree model for use of laparoscopic radical nephrectomy. The odds ratio associated with L_1 is 2.1 (95% CI 1.7–2.6), that with L_2 is 0.38 (95% CI 0.29–0.48), and that with L_3 is 0.29 (95% CI 0.23–0.38).

new technologies and potentially reflect role of resources and access to informational externalities that help promote the adoption of these technologies.

According to Donabedian's structure-process-outcome model for quality-of-care assessment, characteristics of individual providers and their practice environments are structural measures that influence patient outcomes both directly and through their influence on specific processes of care [23]. In fact, these links have been validated empirically in multiple, diverse clinical settings. One well-characterized example is the inverse association between surgeon case volume (a provider characteristic and structural measure) and operative mortality (a patient outcome) following high-risk cancer surgery [24]. Likewise, among patients with prostate cancer, evidence-based utilization of androgen deprivation therapy (a process of care) varies based on characteristics of the treating urologist, including years since medical school graduation and academic affiliation [22]. In addition to a surgeon's individual characteristics, the practice environment also influences treatment decisions and patient outcomes. For instance, patients receiving care at the National Cancer Institute- (NCI-) designated cancer centers have lower adjusted mortality rates following surgical resection of gastric, lung, colorectal, and esophageal cancers than in non-NCI-designated hospitals [25]. Specific to urology, patients treated by physicians in solo practice receive less-frequent surveillance (a process of care) following a bladder cancer diagnosis than do those whose surgeon is in a group practice [26].

Our results are in keeping with existing literature that describes the influence of provider characteristics and practice environments on the adoption of innovative surgical therapies. For example, prior work identified younger surgeon age, active board certification, urban practice location, group practice affiliation, and a competitive practice setting as important facilitators of general surgeons' adoption of laparoscopic cholecystectomy [27, 28]. Similar findings have been described for surgical treatment in early stage breast cancer [29, 30], as well as urological cancers, such as the use of partial nephrectomy for kidney cancer [31], utilization of continent reconstruction among patients undergoing radical cystectomy for bladder cancer [32], and use of androgen deprivation therapy among patients with localized prostate cancer [22].

This study has several limitations. Because SEER-Medicare data are limited to patients >65 years of age, our findings may not apply to younger patients with kidney cancer. Second, similar to surgery for early-stage breast cancer, clarification of the optimal use of partial nephrectomy and laparoscopy will require a better understanding of patient attitudes and preferences that cannot be assessed using claims data. Third, as we used Medicare claims, we may be underestimating the operative volume of individual surgeons treating patients younger than 65 years. Fourth, we could measure only a limited set of surgeon and practice environment characteristics (most of which are structural in nature); as such, there is a need for future studies that assess the degree to which difficult-to-measure barriers such as technical complexity and/or an absence of adopters in their local communities influence urologists' uptake of these newer surgical therapies.

These limitations notwithstanding, our findings have implications for efforts aimed at facilitating the adoption of partial nephrectomy and laparoscopic radical nephrectomy. As described previously, renewed efforts are needed to better understand barriers to initial and sustained adoption among urologists working in rural environments, small practice settings, and those not affiliated with academic medical centers and/or NCI-designated cancer centers. Although more

recently trained urologists were more likely to use laparoscopic radical nephrectomy, our findings counter the notion that uniform adoption will occur naturally as training in this minimally invasive technique becomes more commonplace. Recognizing that social connections and local informational resources facilitate the diffusion of new surgical therapies [27, 33, 34], we see innovative collaborations between urologists, informed by established practice-based surgical research models [35, 36], as representing a potential mechanism for accelerating uniform and equitable adoption of these newer technologies. That being said, the most significant implications from the current study relate to our illustration, more generally, of the power of logic regression as a novel method for discovering interactions in health services and outcomes research. In addition to characteristics of the surgeon and practice environment, others have described multiple contextual factors that influence technology adoption, including, among others, patient demand, professional impact (i.e., financial and social costs), commercial promotion, and magnitude of perceived clinical benefit [37]. As such, methods that allow better characterization and understanding of the complex interplay between these factors will undoubtedly facilitate targeted and efficient interventions to optimize the adoption of both beneficial and potentially harmful new technologies.

Conflict of Interests

The authors declare that they have no financial or nonfinancial competing interests.

Authors' Contribution

Mousumi Banerjee conceived the study, participated in its design, statistical analysis, and interpretation of data, and drafted the paper. Christopher Filson participated in acquisition of data, interpretation of results, and critical review of the paper. Rong Xia participated in statistical analysis and interpretation of data. David C. Miller participated in the study concept, acquisition of data, and interpretation of results and helped to draft the paper. All authors read and approved the final paper.

Acknowledgment

Dr. Banerjee's research was supported by Grant P30-CA46592-05 from the NCI.

References

[1] A. C. Novick, "Laparoscopic and partial nephrectomy," *Clinical Cancer Research*, vol. 10, pp. 6322S–6327S, 2004.

[2] D. C. Miller, J. T. Wei, R. L. Dunn, and B. K. Hollenbeck, "Trends in the diffusion of laparoscopic nephrectomy," *Journal of the American Medical Association*, vol. 295, no. 21, pp. 2480–2482, 2006.

[3] D. C. Miller, D. A. Taub, R. L. Dunn, J. T. Wei, and B. K. Hollenbeck, "Laparoscopy for renal cell carcinoma: diffusion versus regionalization?" *The Journal of Urology*, vol. 176, no. 3, pp. 1102–1107, 2006.

[4] A. F. Fergany, K. S. Hafez, and A. C. Novick, "Long-term results of nephron sparing surgery for localized renal cell carcinoma: 10-year followup," *The Journal of Urology*, vol. 163, no. 2, pp. 442–445, 2000.

[5] A. S. Go, G. M. Chertow, D. Fan, C. E. McCulloch, and C.-Y. Hsu, "Chronic kidney disease and the risks of death, cardiovascular events, and hospitalization," *The New England Journal of Medicine*, vol. 351, no. 13, pp. 1296–1305, 2004.

[6] W. C. Huang, A. S. Levey, A. M. Serio et al., "Chronic kidney disease after nephrectomy in patients with renal cortical tumours: a retrospective cohort study," *The Lancet Oncology*, vol. 7, no. 9, pp. 735–740, 2006.

[7] J. M. Hollingsworth, D. C. Miller, S. Daignault, and B. K. Hollenbeck, "Rising incidence of small renal masses: a need to reassess treatment effect," *Journal of the National Cancer Institute*, vol. 98, no. 18, pp. 1331–1334, 2006.

[8] B. K. Hollenbeck, D. A. Taub, D. C. Miller, R. L. Dunn, and J. T. Wei, "National utilization trends of partial nephrectomy for renal cell carcinoma: a case of underutilization?" *Urology*, vol. 67, no. 2, pp. 254–259, 2006.

[9] D. C. Miller, C. S. Saigal, M. Banerjee, J. Hanley, and M. S. Litwin, "Diffusion of surgical innovation among patients with kidney cancer," *Cancer*, vol. 112, no. 8, pp. 1708–1717, 2008.

[10] I. Ruczinski, C. Kooperberg, and M. Leblanc, "Logic regression," *Journal of Computational and Graphical Statistics*, vol. 12, no. 3, pp. 475–511, 2003.

[11] I. Ruczinski, C. Kooperberg, and M. L. LeBlanc, "Exploring interactions in high-dimensional genomic data: an overview of logic regression, with applications," *Journal of Multivariate Analysis*, vol. 90, no. 1, pp. 178–195, 2004.

[12] C. Kooperberg, J. C. Bis, K. D. Marciante, S. R. Heckbert, T. Lumley, and B. M. Psaty, "Logic regression for analysis of the association between genetic variation in the renin-angiotensin system and myocardial infarction or stroke," *American Journal of Epidemiology*, vol. 165, no. 3, pp. 334–343, 2007.

[13] M. Banerjee and A. Noone, "Tree-based methods for survival data," in *Statistical Advances in the Biomedical Sciences*, A. Biswas, S. Datta, J. P. Fine, and M. R. Segal, Eds., pp. 265–285, John Wiley & Sons, Hoboken, NJ, USA, 2008.

[14] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[15] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth, Belmont, Calif, USA, 1984.

[16] P. J. M. Laarhoven van and E. H. L. Aarts, *Simulated Annealing: Theory and Applications*, Kluwer Academic, Norwell, Mass, USA, 1987.

[17] J. L. Warren, C. N. Klabunde, D. Schrag, P. B. Bach, and G. F. Riley, "Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population," *Medical Care*, vol. 40, no. 8, pp. IV-3–IV-18, 2002.

[18] A. V. Diez Roux, S. S. Merkin, D. Arnett et al., "Neighborhood of residence and incidence of coronary heart disease," *The New England Journal of Medicine*, vol. 345, no. 2, pp. 99–106, 2001.

[19] M. Charlson, T. P. Szatrowski, J. Peterson, and J. Gold, "Validation of a combined comorbidity index," *Journal of Clinical Epidemiology*, vol. 47, no. 11, pp. 1245–1251, 1994.

[20] R. A. Deyo, D. C. Cherkin, and M. A. Ciol, "Adapting a clinical comorbidity index for use with ICD-9-CM administrative

databases," *Journal of Clinical Epidemiology*, vol. 45, no. 6, pp. 613–619, 1992.

[21] R. Morrill, J. Cromartie, and G. Hart, "Metropolitan, urban, and rural commuting areas: toward a better depiction of the United States settlement system," *Urban Geography*, vol. 20, no. 8, pp. 727–748, 1999.

[22] V. B. Shahinian, Y.-F. Kuo, J. L. Freeman, E. Orihuela, and J. S. Goodwin, "Characteristics of urologists predict the use of androgen deprivation therapy for prostate cancer," *Journal of Clinical Oncology*, vol. 25, no. 34, pp. 5359–5365, 2007.

[23] A. Donabedian, "The quality of care. How can it be assessed?" *Journal of the American Medical Association*, vol. 260, no. 12, pp. 1743–1748, 1988.

[24] J. D. Birkmeyer, T. A. Stukel, A. E. Siewers, P. P. Goodney, D. E. Wennberg, and F. L. Lucas, "Surgeon volume and operative mortality in the United States," *The New England Journal of Medicine*, vol. 349, no. 22, pp. 2117–2127, 2003.

[25] N. J. O. Birkmeyer, P. P. Goodney, T. A. Stukel, B. E. Hillner, and J. D. Birkmeyer, "Do cancer centers designated by the National Cancer Institute have better surgical outcomes?" *Cancer*, vol. 103, no. 3, pp. 435–441, 2005.

[26] D. Schrag, L. J. Hsieh, F. Rabbani, P. B. Bach, H. Herr, and C. B. Begg, "Adherence to surveillance among patients with superficial bladder cancer," *Journal of the National Cancer Institute*, vol. 95, no. 8, pp. 588–597, 2003.

[27] J. J. Escarce, "Externalities in hospitals and physician adoption of a new surgical technology: an exploratory analysis," *Journal of Health Economics*, vol. 15, no. 6, pp. 715–734, 1996.

[28] J. J. Escarce, B. S. Bloom, A. L. Hillman, J. A. Shea, and J. S. Schwartz, "Diffusion of laparoscopic cholecystectomy among general surgeons in the United States," *Medical Care*, vol. 33, no. 3, pp. 256–271, 1995.

[29] K. A. Vanderveen, D. A. Paterniti, R. L. Kravitz, and R. J. Bold, "Diffusion of surgical techniques in early stage breast cancer: variables related to adoption and implementation of sentinel lymph node biopsy," *Annals of Surgical Oncology*, vol. 14, no. 5, pp. 1662–1669, 2007.

[30] S. T. Hawley, T. P. Hofer, N. K. Janz et al., "Correlates of between-surgeon variation in breast cancer treatments," *Medical Care*, vol. 44, no. 7, pp. 609–616, 2006.

[31] D. C. Miller, S. Daignault, J. S. Wolf et al., "Hospital characteristics and use of innovative surgical therapies among patients with kidney cancer," *Medical Care*, vol. 46, no. 4, pp. 372–379, 2008.

[32] J. L. Gore, C. S. Saigal, J. M. Hanley, M. Schonlau, and M. S. Litwin, "Variations in reconstruction after radical cystectomy," *Cancer*, vol. 107, no. 4, pp. 729–737, 2006.

[33] E. M. Rogers, *Diffusion of Innovations*, Free Press, New York, NY, USA, 2003.

[34] J. A. Myers and A. Doolas, "How to teach an old dog new tricks and how to teach a new dog old tricks: bridging the generation gap to push the envelope of advanced laparoscopy," *Surgical Endoscopy and Other Interventional Techniques*, vol. 20, no. 8, pp. 1177–1178, 2006.

[35] N. J. O. Birkmeyer and J. D. Birkmeyer, "Strategies for improving surgical quality—should payers reward excellence or effort?" *The New England journal of medicine*, vol. 354, no. 8, pp. 864–870, 2006.

[36] J. M. Westfall, J. Mold, and L. Fagnan, "Practice-based research—"Blue highways" on the NIH roadmap," *Journal of the American Medical Association*, vol. 297, no. 4, pp. 403–406, 2007.

[37] C. B. Wilson, "Adoption of new surgical technology," *British Medical Journal*, vol. 332, no. 7533, pp. 112–114, 2006.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 503132, 7 pages
http://dx.doi.org/10.1155/2014/503132

Research Article

A Two-Stage Exon Recognition Model Based on Synergetic Neural Network

Zhehuang Huang[1,2] and Yidong Chen[2,3]

[1] *School of Mathematics Sciences, Huaqiao University, Quanzhou 362021, China*
[2] *Cognitive Science Department, Xiamen University, Xiamen 361005, China*
[3] *Fujian Key Laboratory of the Brain-Like Intelligent Systems, Xiamen 361005, China*

Correspondence should be addressed to Yidong Chen; ydchen@xmu.edu.cn

Received 30 January 2014; Revised 27 February 2014; Accepted 3 March 2014; Published 25 March 2014

Academic Editor: Xiao-Qin Xia

Exon recognition is a fundamental task in bioinformatics to identify the exons of DNA sequence. Currently, exon recognition algorithms based on digital signal processing techniques have been widely used. Unfortunately, these methods require many calculations, resulting in low recognition efficiency. In order to overcome this limitation, a two-stage exon recognition model is proposed and implemented in this paper. There are three main works. Firstly, we use synergetic neural network to rapidly determine initial exon intervals. Secondly, adaptive sliding window is used to accurately discriminate the final exon intervals. Finally, parameter optimization based on artificial fish swarm algorithm is used to determine different species thresholds and corresponding adjustment parameters of adaptive windows. Experimental results show that the proposed model has better performance for exon recognition and provides a practical solution and a promising future for other recognition tasks.

1. Introduction

With the completion of human genome project, gene data increase exponentially. Identifying the genes encoding of DNA [1] has important theoretical and practical implications. How to quickly access accurate genetic information is an urgent problem to be solved.

Early exon recognition methods were based mainly on statistical models [2], which get their chromosomal order by statistical analysis of different genes. But with the increase of genomic number, statistical methods cannot meet the need for rapid recognition of exons. At present, exon recognition methods based on digital signal processing have also been widely used [3–5]. These techniques select a suitable mapping method and transformation method to get spectral values and identify exons according to fixed length window. Limitations of these methods include slow recognition speed and inability to accurately determine the threshold for different species.

Synergetic theory [6] is the science proposed by Haken to describe high dimension and nonlinear problem as a set of low-dimension nonlinear equations. One advantage of synergetic neural network is that the method is robust against noise and the method can better handle the fuzzy matching problem [7–9]. Exon recognition can also be considered a problem of pattern recognition, for which the proposed method can be used to solve.

Artificial fish swarm algorithm (AFSA) [10, 11] is a class of swarm intelligence optimization algorithms based on the behavior of animals proposed in 2002; the basic idea of AFSA is to imitate the fish behaviors such as praying, swarming, and following. AFSA is very suitable for solving a variety of numerical optimization problems, making the algorithm become a hot topic in the current optimization field quickly. Because of simplicity in principle and good robustness, AFSA has been applied successfully to all kinds of optimization problems such as image segmentation [12], color quantization [13], neural network [14], fuzzy logic controller [15], multi-robot task scheduling [16], fault diagnosis in mine hoist [17], data clustering [18], and other areas.

In this paper, we proposed a two-stage exon recognition model based on synergetic neural network and artificial fish

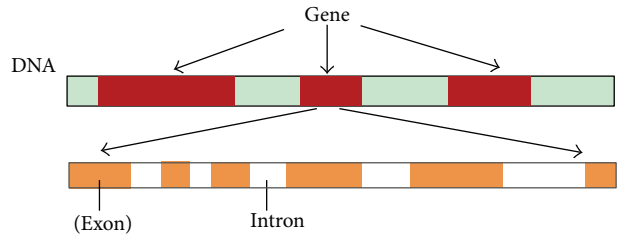


Figure 1: Structure of eukaryotic DNA sequence.

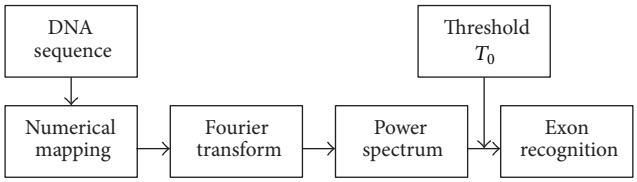


Figure 2: Exon recognition algorithm based on 3-Cycle spectrum.

Table 1

$\Delta x[n]$	1	−1	1	−1	−1	1	1
$\Delta y[n]$	1	1	−1	−1	−1	1	−1
$\Delta z[n]$	1	−1	−1	1	1	1	−1

swarm algorithm. This paper is organized as follows. Firstly, traditional exon recognition method based on digital signal processing and related work are presented. Secondly, an exon recognition model based on synergetic neural network and parameter optimization method based on artificial fish swarm algorithm are introduced. Finally some experimental tests, results, and conclusions are given on the systems.

2. Introduction to Exon Recognition Method Based on Digital Signal Processing

The gene is usually divided into many fragments. The coding sequence is called exons and noncoding part is called introns, as shown in Figure 1.

The objective of gene recognition is to identify the exons of DNA sequence. Gene recognition based on digital signal processing methods consists of several steps [19, 20]. First, gene sequences are transformed into digital symbol sequences using mapping methods [21–24]. This is followed by calculation of the corresponding frequency value by fast Fourier transform and the 3-Cycle properties of the spectrum are then used to identify exons [25, 26]. Finally, fixed sliding window method is used for automatic exon recognition, as shown in Figure 2.

2.1. Z-Curve Mapping. In order to make digital processing, we must transform four nucleotide sequences A, T, G, and C into their corresponding numeric sequence based on certain rules.

Let the four instruction sequences be $\{u_b[n]\}$, $b \in I = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$, and cumulative sequence b_n $(n = 0, 1 \ldots, N-1)$ is $b_n = \sum_{i=0}^{n-1} u_b[i]$; then we can define three sequences $x[n]$, $y[n]$, and $z[n]$:

$$\begin{aligned} x[n] &= 2(\mathrm{A}_n + \mathrm{G}_n) - n, \\ y[n] &= 2(\mathrm{A}_n + \mathrm{C}_n) - n, \\ z[n] &= 2(\mathrm{A}_n + \mathrm{T}_n) - n. \end{aligned} \tag{1}$$

Let

$$\begin{aligned} x[-1] = 0, \quad y[-1] = 0, \quad z[-1] = 0, \\ \Delta x[n] = x[n] - x[n-1], \quad \Delta y[n] = y[n] - y[n-1], \\ \Delta z[n] = z[n] - z[n-1]. \end{aligned} \tag{2}$$

Thus we can get the Z-curve mapping:

$$\begin{pmatrix} \Delta x[n] \\ \Delta y[n] \\ \Delta z[n] \end{pmatrix} = \begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} u_{\mathrm{A}}[n] \\ u_{\mathrm{C}}[n] \\ u_{\mathrm{G}}[n] \\ u_{\mathrm{T}}[n] \end{pmatrix}. \tag{3}$$

For example, the DNA sequence of $S(n)$ is ACGTTAG; then the corresponding Z-curve mapping sequence is shown in Table 1.

2.2. The Power Spectrum. To study the characteristics of DNA coding sequences (exons), we can do the discrete Fourier transform (DFT), respectively, for the instruction sequences:

$$U_b[k] = \sum_{n=0}^{N-1} u_b[n] e^{-j(2\pi nk/N)}, \quad k = 0, 1, \ldots, N-1. \tag{4}$$

Thus we can calculate the power spectrum:

$$\begin{aligned} P_z[k] = |\Delta X[k]|^2 + ||\Delta Y[k]||^2 + |\Delta Z[k]|^2, \\ k = 0, 1, \ldots, N-1, \end{aligned} \tag{5}$$

where $\Delta X[k]$, $\Delta Y[k]$, and $\Delta Z[k]$ are the Fourier transform of $\Delta x[n]$, $\Delta y[n]$, and $\Delta z[n]$, respectively.

The spectral peaks of exon sequences are larger in $k = N/3$ and $k = 2N/3$ of the power spectrum curve, while they are not similar for intron. This statistical phenomenon is known as 3-Cycle. Suppose that the average power spectrum of DNA sequences is

$$\overline{E} = \frac{\sum_{k=0}^{N-1} P[k]}{N}. \tag{6}$$

The power spectrum ratio of the DNA sequence and the average spectrum of the entire sequence are known as SNR (signal-to-noise ratio):

$$R = \frac{P[N/3]}{\overline{E}}. \tag{7}$$

Figure 3 shows the power spectrum of viral genes.

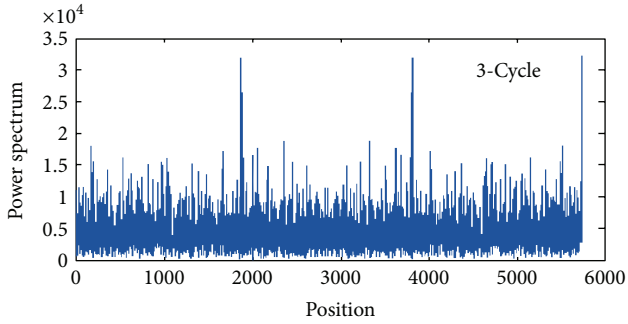


FIGURE 3: The power spectrum of viral gene sequence.

From Figure 3, we can see that the spectrum presents obvious 3-Cycle. The peaks appear roughly in 2000, 4000, and 6000. So the exon segment can be determined, enabling the recognition of genes.

The highest point of power spectrum may not appear in $k = N/3$ and $k = 2N/3$ but occur in the surrounding. So we can calculate average SNR R_1 and R_2 of intervals $[N/3 - \gamma, N/3 + \gamma]$ and $[2N/3 - \gamma, 2N/3 + \gamma]$, respectively:

$$R_1 = \frac{\sum_{k=N/3-\gamma}^{N/3+\gamma} P[E]}{(2\gamma+1)\,\overline{E}}, \qquad R_2 = \frac{\sum_{k=2N/3-\gamma}^{2N/3+\gamma} P[E]}{(2\gamma+1)\,\overline{E}}. \tag{8}$$

2.3. Automatic Recognition Algorithm Based on Fixed Sliding Windows. Supposed M is the length of fixed window; we can do four discrete Fourier transforms (DFT) for instruction sequences $\{u_b[n]\}$ $(0 \le n \le N-1)$,

$$U_b[k] = \sum_{i=n-(M-1)/2}^{i=n+(M-1)/2} u_b[i]\, e^{-j(2\pi ik/M)}, \quad k = 0, 1, \ldots, M-1. \tag{9}$$

Then the total spectrum $p(n; M/3)$ at position $M/3$ is

$$\begin{aligned} P\left[\frac{M}{3}\right] &= \left|U_A\left[\frac{M}{3}\right]\right|^2 + \left|U_T\left[\frac{M}{3}\right]\right|^2 \\ &\quad + \left|U_G\left[\frac{M}{3}\right]\right|^2 + \left|U_C\left[\frac{M}{3}\right]\right|^2 \\ &\triangleq p\left(n; \frac{M}{3}\right). \end{aligned} \tag{10}$$

3. Related Work

The SNR of exon sequences reflects the distribution of spectrum peak. SNR greater than a given threshold is a characteristic of exons, while introns generally do not have this property.

Protein coding regions and noncoding regions can be distinguished using the value of SNR, but this method still has a large predictive error because the spectrum peak varies amongst different biological categories. A fixed threshold is unreasonable to use for different biological categories. Therefore, determining the SNR threshold has great significance for exon recognition. Note that it is difficult to find the proper prediction threshold for biological categories when relying only on prior biological knowledge.

Xu [27] proposed a method based on bootstrap algorithm to determine the best SNR threshold that can be obtained from marked exon sequences. The results of that study showed that the average prediction accuracy of the method was 81%, which is 19% higher than other methods that employ empirical thresholds. In paper [28], a novel model was proposed to determine the SNR threshold based on the means of biological categories and improved the recognition performance to some extent.

But all the methods mentioned above have problems, such as slow recognition speed, inaccurate determination of the threshold for different species, and the requirement to know the exon fragments of DNA sequences. In the following sections, we propose a novel two-stage exon recognition model based on synergetic neural network and artificial fish swarm algorithm to better deal with these problems.

4. A Novel Two-Stage Exon Recognition Model

In this section, a two-stage exon recognition model is presented. In the first stage, synergetic neural network is used to determine initial exon intervals. In the second stage, final accurate exon intervals determination based on adaptive sliding window and parameter optimization algorithm are introduced.

4.1. Initial Exon Intervals Determination Based on Synergetic Neural Network. The basic principle of synergetic neural network [29, 30] is that the pattern recognition procedure can be viewed as the competition progress of many order parameters. The strongest order parameter will win by competition and desired pattern will be recognized.

A pattern that remained to be recognized, q, is constructed by a dynamic process which translates q into one of prototype pattern vectors v_k through status $q(t)$; namely, this prototype pattern is closest to $q(0)$. The process is described as the following equation:

$$q \longrightarrow q(t) \longrightarrow v_k. \tag{11}$$

A dynamic equation can be given for an unrecognized pattern q:

$$\begin{aligned} \dot{q} &= \sum_{k=1}^{M} \lambda_k v_k \left(v_k^+ q\right) - B \sum_{k' \neq k} \left(v_{k'}^+ q\right)^2 \left(v_k^+ q\right) v_k \\ &\quad - C\left(q^+ q\right) q + F(t), \end{aligned} \tag{12}$$

where q is the status vector of input pattern with initial value q_0, λ_k is attention parameter, v_k is prototype pattern vector, and v_k^+ is the adjoint vector of v_k that satisfies

$$\left(v_k^+, v_{k'}^T\right) = v_k^+ \cdot v_{k'}^T = \delta_{kk'}. \tag{13}$$

Table 2: The signal-to-noise ratio of four different gene sequences.

Gene categories	Exon			Intron		
	Number	R-mean	Variance	Number	R-mean	Variance
Human	35	3.02	3.071	26	0.82	0.533
Mus musculus	357	2.46	2.508	275	0.68	0.414
Sewer rat	45	3	5.233	35	0.83	0.624
Mammalian	827	2.72	6.243	626	0.67	0.394

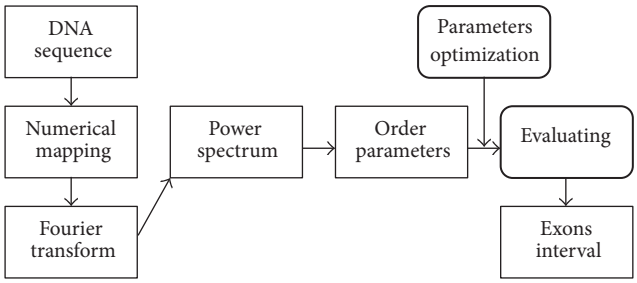


Figure 4: Exon recognition based on synergetic neural network.

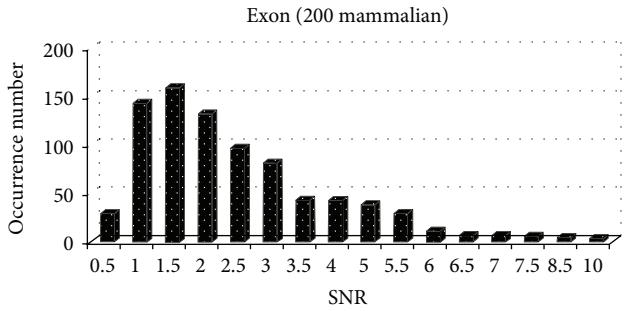


Figure 5: The SNR distribution of 200 mammalian exons.

Corresponding dynamic equation of order parameters is

$$\dot{\xi}_k = \lambda_k \xi_k - B \sum_{k' \neq k} \xi_{k'}^2 \xi_k - \mathrm{C} \left| \sum_{k'=1}^{M} \xi_{k'}^2 \right| \xi_k. \tag{14}$$

Haken has proved that when $\lambda_k = c$ $(c > 0)$, the largest initial order parameter will win and the network will then converge.

We firstly introduce the synergetic theory to exon recognition; an exon recognition algorithm based on synergetic neural network is shown in Figure 4.

We use synergetic neural network and N equal method to quickly determine the initial exon region, as shown in Algorithm 1.

4.2. Get Precise Exon Intervals Using Adaptive Smoothing Window. We can obtain several possible exon intervals by Algorithm 1. In this section, we propose an adaptive sliding window algorithm to get more accurate intervals, as shown in Algorithm 2.

4.3. Parameter Optimization Based on Artificial Fish Swarm Algorithm. The parameters T_0 and γ directly influence the performance of exon recognition. The adjustment of the parameters is a global behaviour and has no general research theory to control the parameters in the recognition process at present. In this section, artificial fish swarm algorithm is used to search the global optimum parameters (T_0, γ) in the corresponding parameter space.

The parameter optimization based on artificial fish swarm algorithm is shown as Algorithm 3.

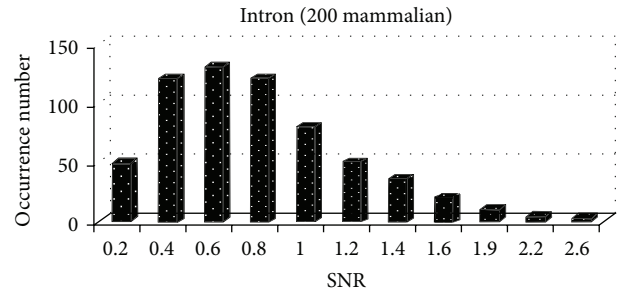


Figure 6: The SNR distribution of 200 mammalian introns.

5. Experiment

5.1. Data Description. In our experiments, we use some gene sequences provided by Chinese Graduate Mathematical Contest in Modeling. Chinese graduate Mathematical Contest in modeling is aimed at improving the students' comprehensive abilities of mathematical modeling and computer to solve practical problems. From different points of view, the integrated use of a variety of mathematical methods established the mathematical model of the characteristic.

We selected 100 human gene sequences, 100 rodent gene sequences (including *Mus musculus* and Sewer rat), and 200 mammalian gene sequences for testing. The signal-to-noise ratios of the sequences are gotten by SPSS statistical analysis software, as shown in Table 2.

From Table 2, we can find out that the difference between SNR standard deviation of exons is greater than SNR standard deviation of introns.

At the same time, we analyze the SNR distribution of exons and introns of 200 mammalian gene sequences, as shown in Figure 5 and Figure 6.

From Figure 5 and Figure 6, we can see that the mammalian introns are mostly less than 2, while exons are mostly distributed in the range of $[0, 2]$, which accounts for 55.38%. Therefore, it is unreasonable to set SNR threshold of different categories as fixed value. How to accurately determine SNR

(1) Let S is a given gene sequence, S_{start} and S_{end} are the beginning and end of the sequence respectively, T_0 is throdthod of spectral values;
(2) Using Z_Curve mapping converted gene sequence to the corresponding numeric sequence;
(3) Using fast Fourier transform to get spectral values R_1 and R_2 according to the formula (8);
(4) Calculating gene sequence order parameter:

$$\xi_1 = \frac{R_1}{R_1 + R_2}, \qquad \xi_2 = \frac{R_2}{R_1 + R_2};$$

(5) Setting network parameter λ_k and B, C;
(6) Order parameter evolution according to formula (14);
(7) If $\xi_1 > T_0$ and $\xi_2 > T_0$, then $[S_{\text{start}}, S_{\text{end}}]$ is recorded as a possible interal, and S is divided equally into n interals $S_1, S_2, \ldots, S_n$, Repeat step 1 to step 7;
(8) End.

Algorithm 1: Determination of initial exon region based on synergetic neural network.

(1) Let W is a given gene sequence, W_{start} and W_{end} are the beginning and the end of the sequence respectively;
(2) Using Z_Curve mapping converted gene sequence $[W_{\text{start}}, W_{\text{end}}]$ to the corresponding numeric sequence;
(3) Using fast Fourier transform to get spectral values;
(4) Calculating gene sequence order parameter:

$$\xi_1 = \frac{R_1}{R_1 + R_2}, \qquad \xi_2 = \frac{R_2}{R_1 + R_2};$$

(5) Order parameter evolution according to formula (14);
(6) If $\xi_1 > T_0$, $\xi_2 > T_0$ and $W_{\text{start}} + \gamma < W_{\text{end}} - \gamma$, Then $W_{\text{start}} = W_{\text{start}} + \gamma$, $W_{\text{end}} = W_{\text{end}} - \gamma$,
Repeat step 2 to step 6;
(7) Output the final interval $[W_{\text{start}}, W_{\text{end}}]$.

Algorithm 2: Precise exon regions based on adaptive smoothing window.

threshold of each kind of biological gene has important significance.

5.2. Experiment Results. Suppose that sensitivity $S_N = T_P/(T_p + F_N)$ and specificity $S_P = T_N/(T_N + F_P)$, where T_P is the number of exons which are correctly identified, T_N is the number of introns which are correctly identified, F_P is the number of exons which are not correctly identified, and F_N is the number of introns which are not correctly identified. Then we can compute the accurate rate $A_c = (S_N + S_P)/2$.

For comparison, we use four strategies.

(1) Baseline: automatic recognition algorithm with threshold $R_0 = 2$.
(2) Bootstrap: the threshold selection algorithm based on bootstrap method.
(3) SNN: exon recognition based on synergetic neural network.
(4) SNN + AFSA: two-stage exon recognition model based on synergetic neural network and artificial fish swarm algorithm.

The testing performance of Baseline is shown as in Table 3.

The experiments showed that when the exon length is short, the recognition accuracy rate is low. In the short gene coding sequence, 3-base periodicity is not absolutely satisfied. In our experiments, we complete a two-stage exon recognition model based on synergetic neural network and artificial fish swarm algorithm. The parameter settings of artificial fish swarm algorithm are shown in Table 4.

In the experiment, we set the recognition accuracy rate as score function.

The testing performance of SNN + AFSA is shown as in Table 5.

Table 5 shows that the two-stage exon recognition algorithm improves precision compared to the Baseline system. Experiments also indicate that the improved model has a more powerful global exploration ability and a reasonable convergence speed.

The accurate rate A_c of different methods is shown in Table 6.

Detailed comparisons of results are given in Table 6. Experimental results show that the proposed model SNN and SNN + AFSA have good performance for exon recognition. The accurate rate we obtained for all four corpuses is comparable to the state-of-the-art systems, such as Baseline and bootstrap method. Through the evaluating of order parameter equation of SNN to obtain the best threshold, we can further improve the exon recognition performance.

At the same time, we can see that the performance of SNN + AFSA is better than SNN model. This is because

(1) Initialize the parameters of artificial fish, such as *step*, *visual*, the number of exploratory, maximum number of iterations, and randomly generated n fishes;
(2) Set bulletin board to record the current status of each fish, and select the optimal value;
(3) Implementation of prey behavior, swarm behavior and follow behavior;
(4) Optimal value in bulletin board is updated;
(5) If termination condition is satisfied, output the result; otherwise return to step 2.

Algorithm 3: Parameter optimization based on artificial fish swarm algorithm.

Table 3: The testing performance of Baseline.

Gene categories	T_P	F_N	S_N	T_N	F_P	S_P	A_c
Human	17	18	0.485	24	2	0.923	0.71
Mus musculus	146	211	0.409	271	4	0.985	0.70
Sewer rat	17	28	0.378	31	4	0.886	0.63
Mammalian	369	458	0.446	621	5	0.992	0.72

Table 4: The parameter settings of artificial fish swarm algorithm.

Algorithm	Fish number	Visual	Delta	Step	Number of iterations
AFSA	100	2.85	9	1	60

Table 5: The test performance of SNN + AFSA.

Gene categories	T_P	F_N	S_N	T_N	F_P	S_P	A_c
Human	30	5	0.857	19	7	0.731	0.79
Mus musculus	295	62	0.826	220	55	0.80	0.81
Sewer rat	36	9	0.80	28	7	0.80	0.80
Mammalian	630	197	0.762	607	19	0.97	0.87

Table 6: The test performance comparison among different methods.

Gene categories	Baseline	Bootstrap	SNN	SNN + AFSA
Human	0.71	0.76	0.78	0.79
Mus musculus	0.70	0.78	0.80	0.81
Sewer rat	0.63	0.75	0.77	0.80
Mammalian	0.72	0.84	0.85	0.87

the attention parameters are very important for SNN and optimization algorithm is essential for better performance. Experimental results show that improved AFSA algorithm has better global and local parameter searching capabilities and thus a better recognition result.

It is worth noting that experimental results show that run times of our proposed model reduced with good speedup ratio compared with Baseline. Further studies show that the procedure exhibits data parallelism, so it can be effectively parallelized by running it concurrently. In the future work, we will utilize parallel processing techniques for rapid exon recognition based on SNN to further reduce the run time.

6. Conclusions

In the paper, we proposed a two-stage exon recognition model based on synergetic neural network and artificial fish swarm algorithm. Experiments show that the proposed model can improve the precision of exon recognition.

We got the following conclusions.

(1) The exon recognition procedure can be viewed as the competition progress of many order parameters. The proposed model based on synergetic neural network and N equal method can quickly determine the exon intervals.

(2) Artificial fish swarm algorithm has both global and local search ability and can effectively choose the parameters of our proposed model.

(3) Using N equal algorithm to obtain exon intervals may still miss some intervals which are in the middle; we will further improve the algorithm or use different pattern recognition algorithm in the future.

It must be noted that, although we have made some efforts to explore the intelligent exon recognition algorithm in this paper. But due to the special nature of life science itself, there are many problems such as how to accurately determine that the exon interval needs further study. But we believe that with the development of social progress and technology, gene identification technology will become increasingly perfect; we expect it can bring gospel to manking in the near future.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant no. 61005052), the Fundamental Research Funds for the Central Universities (Grant no. 2010121068), the Natural Science Foundation of Fujian Province of China (Grant no. 2011J01369), and the Science and Technology Project of Quanzhou (Grant no. 2012Z91).

References

[1] C. B. Burge and S. Karlin, "Finding the genes in genomic DNA," *Current Opinion in Structural Biology*, vol. 8, no. 3, pp. 346–354, 1998.

[2] Z. Wang, Y. Chen, and Y. Li, "A brief review of computational gene prediction methods," *Genomics Proteomics Bioinformatics*, vol. 2, no. 4, pp. 216–221, 2004.

[3] S. D. Sharma, K. Shakya, and S. N. Sharma, "Evaluation of DNA mapping schemes for exon detection," in *Proceedings of the International Conference on Computer, Communication and Electrical Technology (ICCCET '11)*, pp. 71–74, March 2011.

[4] D. Kotlar and Y. Lavner, "Gene prediction by spectral rotation measure: a new method for identifying protein-coding regions," *Genome Research*, vol. 13, no. 8, pp. 1930–1937, 2003.

[5] C. Yin and S. S.-T. Yau, "Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence," *Journal of Theoretical Biology*, vol. 247, no. 4, pp. 687–694, 2007.

[6] H. Haken, *Synergetic Computers and Cognition-A Top-Down Approach to Neural Nets*, Springer, Berlin, Germany, 1991.

[7] J. Shao, J. Gao, and X. Z. Yang, "Synergetic face recognition algorithm based on ICA," in *Proceedings of the International Conference on Neural Networks and Brain*, vol. 1, pp. 249–253, Beijing, China, October 2005.

[8] Z. Jiang and R. A. Dougal, "Synergetic contro1 of power converters for pulse current charging of advanced batteries from a fuel cell power source," *IEEE Transactions on Power Electronics*, vol. 19, no. 4, pp. 1140–1150, 2004.

[9] X. L. Ma and L . C. Jiao, "Reconstruction of order parameters based on immunity clonal strategy for image classification," in *Image Analysis and Recognition*, A. Campilho and M. Kamel, Eds., vol. 3211 of *Lecture Notes in Computer Science*, pp. 455–462, Springer, Berlin, Germany, 2004.

[10] X. L. Li, S. H. Feng, J. X. Qian, and F. Lu, "Parameter tuning method of robust pID controller based on artificial fish school algorithm," *Chinese Journal of Information and Control*, vol. 33, no. 1, pp. 112–115, 2004.

[11] X. L. Li, F. Lu, G. H. Tian, and J. X. Qian, "Applications of artificial fish school algorithm in combinatorial optimization problems," *Chinese Journal of Shandong University: Engineering Science*, vol. 34, no. 5, pp. 64–67, 2004.

[12] W. Tian, Y. Geng, J. Liu, and L. Ai, "Optimal parameter algorithm for image segmentation," in *Proceedings of the 2nd International Conference on Future Information Technology and Management Engineering (FITME '09)*, pp. 179–182, December 2009.

[13] D. Yazdani, H. Nabizadeh, E. M. Kosari, and A. N. Toosi, "Color quantization using modified artificial fish swarm algorithm," in *Proceedings of the International conference Artificial Intelligence*, vol. 7106 of *Lecture Notes in Artificial Intelligence*, pp. 382–391, 2011.

[14] M. Zhang, C. Shao, F. Li, Y. Gan, and J. Sun, "Evolving neural network classifiers and feature subset using artificial fish swarm," in *Proceedings of the IEEE International Conference on Mechatronics and Automation (ICMA '06)*, pp. 1598–1602, Luoyang, China, June 2006.

[15] D. Chen, L. Shao, Z. Zhang, and X. Yu, "An image reconstruction algorithm based on artificial fish-swarm for electrical capacitance tomography system," in *Proceedings of the 6th International Forum on Strategic Technology (IFOST '11)*, pp. 1190–1194, August 2011.

[16] C.-J. Wang and S.-X. Xia, "Application of probabilistic causal-effect model based artificial fish-swarm algorithm for fault diagnosis in mine hoist," *Journal of Software*, vol. 5, no. 5, pp. 474–481, 2010.

[17] W. Tian, Y. Tian, L. Ai, and J. Liu, "A new optimization algorithm for fuzzy set design," in *Proceedings of the International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC '09)*, vol. 2, pp. 431–435, August 2009.

[18] Y. Cheng, M. Jiang, and D. Yuan, "Novel clustering algorithms based on improved artificial fish swarm algorithm," in *Proceedings of the 6th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD '09)*, vol. 3, pp. 141–145, August 2009.

[19] M. J. Berryman, A. Allison, C. R. Wilkinson, and D. Abbott, "Review of signal processing in genetics," *Fluctuation and Noise Letters*, vol. 5, no. 4, pp. 13–35, 2005.

[20] N. von Öhsen, I. Sommer, R. Zimmer, and T. Lengauer, "Arby: automatic protein structure prediction using profile-profile alignment and confidence measures," *Bioinformatics*, vol. 20, no. 14, pp. 2228–2235, 2004.

[21] D. Kotlar and Y. Lavner, "Gene prediction by spectral rotation measure: a new method for identifying protein-coding regions," *Genome Research*, vol. 13, no. 8, pp. 1930–1937, 2003.

[22] C. Yin and S. S.-T. Yau, "Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence," *Journal of Theoretical Biology*, vol. 247, no. 4, pp. 687–694, 2007.

[23] A. Rushdi and J. Tuqan, "Gene identification using the stroke Z sign-curve representation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '06)*, vol. 2, pp. 1024–1027, May 2006.

[24] S. D. Sharma, K. Shakya, and S. N. Sharma, "Evaluation of DNA mapping schemes for exon detection," in *Proceedings of the International Conference on Computer, Communication and Electrical Technology (ICCCET '11)*, pp. 71–74, March 2011.

[25] D. Kotlar and Y. Lavner, "Gene prediction by spectral rotation measure: a new method for identifying protein-coding regions," *Genome Research*, vol. 13, no. 8, pp. 1930–1937, 2003.

[26] L. Bo and K. Ding, "Graphical approach to analyzing DNA sequences," *Journal of Computational Chemistry*, vol. 26, no. 14, pp. 1519–1523, 2005.

[27] S. L. Xu, *Threshold selection of gene prediction Based on Bootstrap algorithm [M.S. thesis]*, University of Electronic Science and Technology, Sichuan, China, 2011.

[28] J. F. Shao, X. H. Yan, and S. Shao, "SNR of DNA sequences mapped by general affine transformations of the indicator sequences," *Journal of Mathematical Biology*, vol. 67, no. 2, pp. 433–451, 2013.

[29] Z. Jiang and R. A. Dougal, "Synergetic contro1 of power converters for pulse current charging of advanced batteries from a fuel cell power source," *IEEE Transactions on Power Electronics*, vol. 19, no. 4, pp. 1140–1150, 2004.

[30] J. Gao, H. Dong, J. Shao, and J. Zhao, "Parameters optimization of synergetic recognition approach," *Chinese Journal of Electronics*, vol. 14, no. 2, pp. 192–197, 2005.