Analysis and Modeling of Quality Improvement on Clinical Fitness Landscapes

A Dissertation Presented

by

Narine Manukyan

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fullfillment of the Requirements
for the Degree of Doctor of Philosophy
Specializing in Computer Science

October, 2014

Accepted by the Faculty of the Graduate College, The University of Vermont, in partial fulfillment of the requirements for the degree of Doctor of Philosophy, specializing in Computer Science.

Dissertation Examination Committee:

Advisor
Margaret J. Eppstein, Ph.D.

Donna M. Rizzo, Ph.D.

Jeffrey S. Buzas, Ph.D.

Chairperson
Peter S. Dodds, Ph.D.

Dean, Graduate College
Cynthia J. Forehand, Ph.D.

Date: May, 2014

Abstract

Widespread unexplained variations in clinical practices and patient outcomes, together with rapidly growing availability of data, suggest major opportunities for improving the quality of medical care. One way that healthcare practitioners try to do that is by participating in organized healthcare quality improvement collaboratives (QICs). In QICs, teams of practitioners from different hospitals exchange information on clinical practices, with the aim of improving health outcomes at their own institutions. However, what works in one hospital may not work in others with different local contexts, due to non-linear interactions among various demographics, treatments, and practices. I.e., the clinical landscape is a complex socio-technical system that is difficult to search. In this dissertation we develop methods for analysis and modeling of complex systems, and apply them to the problem of healthcare improvement.

Searching clinical landscapes is a multi-objective dynamic problem, as hospitals simultaneously optimize for multiple patient outcomes. We first discuss a general method we developed for finding which changes in features may be associated with various changes in outcomes at different points in time with different delays in affect. This method correctly inferred interactions on synthetic data, however the complexity and incompleteness of the real hospital dataset available to us limited the usefulness of this approach.

We then discuss an agent-based model (ABM) of QICs to show that teams comprising individuals from similar institutions outperform those from more diverse institutions, under nearly all conditions, and that this advantage increases with the complexity of the landscape and the level of noise in assessing performance. We present data from a network of real hospitals that provides encouraging evidence of a high degree of similarity in clinical practices among hospitals working together in QIC teams. Based on model outcomes, we propose a secure virtual collaboration system that would allow hospitals to efficiently identify potentially better practices in use at other institutions similar to theirs, without any institutions having to sacrifice the privacy of their own data.

To model the search for quality improvement in clinical fitness landscapes, we need benchmark landscapes with tunable feature interactions. NK landscapes have been the classic benchmarks for modeling landscapes with epistatic interactions, but the ruggedness is only tunable in discrete jumps. Walsh polynomials are more finely tunable than NK landscapes, but are only defined on binary alphabets and, in general, have unknown global maximum and minimum.

We define a different subset of interaction models that we dub as NM landscapes. NM landscapes are shown to have smoothly tunable ruggedness and difficulty and known location and value of global maxima. With additional constraints, we can also determine the location and value of the global minima. The proposed NM landscapes can be used with alphabets of any arity, from binary to real-valued, without changing the complexity of the landscape. NM landscapes are thus useful models for simulating clinical landscapes with binary or real decision variables and varying number of interactions. NM landscapes permit proper normalization of fitnesses so that search results can be fairly averaged over different random landscapes with the same parameters, and fairly compared between landscapes with different parameters.

In future work we plan to use NM landscapes as benchmarks for testing various algorithms that can discover epistatic interactions in real world datasets.

Citations

Material from this dissertation has been published in the following form:

JOURNAL PUBLICATIONS

Chapter 1:
Manukyan, N., Eppstein, M.J., Horbar, J.D., Leahy, K.A., Kenny, M.J., Mukherjee, S., and Rizzo, D.M.. "Exploratory Analysis in Time-Varying data sets: a Healthcare Network Application", *International Journal of Advanced Computer Science*, 3(7):322-329, 2013.

Chapter 2:
Manukyan, N., Eppstein, M.J., and Horbar. J..“Team Learning for Healthcare Quality Improvement”, *IEEE Access*, 1:545-557, 2013.

Chapter 3:
Manukyan, N., Eppstein, M.J., and Buzas, J.S.. "NM Landscapes", in review at *IEEE Transactions on Evolutionary Computation*, submitted May 2, 2014.

CONFERENCE PUBLICATIONS

Manukyan, N., Eppstein, M.J., and Horbar, J.D..“Team Structure and Quality Improvement in Collaborative Environments”, *Proceedings of Collaborative Technologies and Systems* (CTS), pp. 523–529, 2013 International Conference, IEEE 2013.

Manukyan, N., Eppstein, M.J., Horbar, J.D., Leahy, K.A., Kenny, M.J., Mukherjee, S., and Rizzo, D.M.. "Evolutionary Mining for Multivariate Associations in Large Time-Varying data sets: a Healthcare Network Application", *Proceedings of the Genetic and Evolutionary Computation Conference* (GECCO), pp. 1547-1548, 2012.

Manukyan, N., Eppstein, M.J., and Buzas, J.S.. "NM Landscapes: Beyond NK", *Proceedings of the Genetic and Evolutionary Computation Conference* (GECCO), to appear July 2014.

Acknowledgements

There are some events in a person's life that can forever alter their future. Often, these events involve meeting people of great importance who inspire us and lead us to be our better selves. Every person is a by-product of his or her environment and the people who contribute to their development. There have been many wonderful people in my life that helped me to grow professionally, intellectually and personally, some of which I would like to acknowledge with respect to this work. First, I would like to thank my adviser Dr. Eppstein, who inspired me with her work before I even met her in person. She is a great example for all women pursuing a science degrees, as she is living proof that one can have an excellent scholarly record, successful career, family, children and still have time for kung fu, hiking and enjoying all those things. I was fortunate to have the opportunity to work with her. She taught me how to be a researcher, think critically, question everything and dare to look for answers, even for the most challenging problems (including NP-complete ones). Without her help and dedication this work would not have been possible, but more importantly I found a friend and mentor for a life. I would like to thank Dr. Rizzo for introducing me to artificial neural networks. Hers was one of the most important classes I ever took, which forever altered who I wanted to be. She inspired me to enjoy machine learning algorithms by taking an intuitive approach and thinking about applications rather than getting lost in the mechanics of details. She was always there when I needed help convincing others about some point or when I needed some encouragement during data-driven research detours. I would like to thank Dr. Dodds for being a great research enthusiast and having a cool attitude towards academia. I really enjoyed working with him as a member of my committee as well as taking courses with him that inspired me to both critique and enjoy academia. I would also like to thank Dr. Buzas for all his help, especially his readi-

ness to discuss scientific questions when I showed up at his office without appointments. I am grateful for his research that inspired me to ask interesting questions about statistics and mathematics. I would also like to thank Dr. Horbar, Mike Kenny and Kathleen Leahy from Vermont Oxford Network for their support and encouragement throughout the data analysis process.

On a more personal note, I would like to thank my amazing and intelligent husband Brian, who has recently become Dr. Hall. He has provided me with great support and encouragement, but most importantly filled my life with joy. I would like to thank my wonderful parents Kliment and Gohar, who have always challenged me in every good way to be successful and provided me with a carefree life that enabled me to pursue my dreams and ambitions. I would also like to thank my amazing sisters Victoria and Zarine, who bring lots of fun and fashion to my life. Finally, I express my hearty gratitude to all of my friends at the University of Vermont and everywhere in the world, especially Armenia, for their support and love.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation, Scientific Questions and Background

Widespread unexplained variations in clinical practices and patient outcomes, together with growing availability of data and computational power, suggest major opportunities for improving the quality of healthcare. While all agree that there is a need for healthcare improvement, there are a range of views as to how this should be accomplished. In 2001 it was proposed that healthcare is an adaptive complex system and that a systems approach could enable researchers and practitioners to properly model changes in healthcare due to evidence-based treatment and diagnoses by multidisciplinary care management teams [32, 96, 97, 140]. According to [96], healthcare is a system with inherent nonlinearity, unpredictability, behavioral patterns and self-organization where agents and systems can adapt to local contingencies and interactions lead to continually emerging new behaviors. Others have similarly described healthcare as a complex system with emergent behavior [113] and a dynamic socio-technical system [24]. This is not a universally accepted view in the healthcare community (e.g., see [104]). However, interest in using a systems approach for making decisions in healthcare is growing as the clinical environment gets increasingly complex, as recognition of this complexity grows, and as diagnosis and treatments become increasingly evidence-based [141]. For example, the World Health Organization now mandates "Understanding systems and the impact of complexity on patient care" in their curriculum on patient safety [87]. There are many challenges associated with studying complex systems as they are often highly coupled and high dimensional [94]. In highly coupled systems recognizing failure as well as taking appropriate recovery actions can be very hard, as they can often trigger multiple, unanticipated effects elsewhere in the system.

Healthcare systems are heterogenous in terms of the reasons for various outcomes, where the same goal can often be achieved in several ways. For example, in neonatal intensive care a neonatologist can select different valid ventilation strategies to deal with the same condition [124]. On the other hand, the price of error is very high in healthcare due to the intrinsic vulnerability of patients and the value of of human life. For example, the mortality rates of some planned clinical procedures regularly exceed 1%, which is much higher than risks in other domains such as aviation or nuclear power [113]. Thus, in healthcare, the search for improvement is not only about the efficacy of the final results, but also about the safety of patients involved in trials leading up to discoveries of improved treatments and practices.

Multi-institutional randomized controlled trials (RCTs) are considered to be the standard of evidence-based medicine. In this approach, trials are conducted within a group of collaborating institutions and the entire healthcare community can learn from the published outcomes, if the results are statistically determinate. However what works in one hospital might not work in others due to differences in local contexts [142]. A limitation of the multi-institutional RCTs is that they often lack the ability to incorporate local contexts of different hospitals and potential interactions among clinical practices due to the large number of patients and resources required for each study, despite numerous studies showing that such interactions exist ([8, 20, 90, 111, 121, 130]). In addition many RCTs are underpowered leading to inconclusive results [111]. Therefore new methods are necessary for determining when to adopt new practices, while taking into consideration the local contexts, relationships and interdependencies among clinical cultures, treatments, practices, demographics and other varying attributes of hospitals.

As policymakers and health care leaders seek effective strategies for healthcare improvement, a new approach called quality improvement collaboratives (QICs) has become increasingly popular. In QICs, multi-institutional teams share information to identify potentially better practices that are subsequently evaluated in the local contexts of specific institutions without strict statistical criteria for success. Several researchers have provided definitions of QICs [3, 23, 64, 78, 89, 95, 123, 139]. These have been consolidated and summarized by Nadeem et. al. [84], as follows:

> "Our study defines QICs as organized, structured group learning initiatives in which the organizers took the following steps: (1) convened multidisciplinary teams representing different levels of the organization; (2) focused on improving specific provider practices or patient outcomes; (3) included training from experts in a particular practice and/or the quality improvement methods; (4) provided a model for improvement with measurable targets, data collection, and feedback; (5) used multidisciplinary teams in active improvement processes in which they implemented small tests of change or engaged in PDSA activities; and (6) employed structured activities and opportunities for learning

> and cross-site communication (e.g., in-person learning sessions, phone calls, email listserves)" [84, page 359].

The current evidence supporting the effectiveness QICs is positive but limited [119]. While some evidence shows that QICs can change provider practices (for example, patient education and medication management [14, 122, 146]), there is limited evidence for positive patient outcomes as a result of QICs [4, 5, 7, 10, 49, 71, 78, 117, 119, 120]). Nine controlled studies of QICs were examined in a systematic literature review [119], the majority of which used matched controls or administrative data for comparison. Out of the nine studies, two had positive effects on study outcomes (patient pain prevalence and infant mortality rates), two studies found no differences, and the rest were mixed [119].

While the US federal and state governments make significant investments in QICs (e.g., see [84]), there is very little understanding of which aspects of these collaborative efforts are linked to positive patient outcomes or positive professional development [78, 119, 123]. Furthermore, it is not clear what are the best mechanisms for implementing QICs. As more information and data become available on QICs, researchers continue to look for evidence for positive outcomes and ways for improvement of QIC implementations. However, there is little theory to guide the search. In this dissertation we take two computational approaches for evaluating healthcare quality improvement strategies.

One computational approach examines previously documented data (on patient outcomes, healthcare practices, social networks of multi-institutional teams seeking to improve healthcare) and tries to use evolutionary data mining techniques for identifying interesting patterns and inferring positive or negative trends. This approach depends on the accuracy and completeness of data collection as well as the methods used for analysis. It is designed for data rich environments, however much of the existing data was not collected with these analysis in mind and it is often not possible to capture all relevant information (for example, how often healthcare professionals interact outside of formal hospital collaboration settings). One of the challenges in this kind of data analysis is that healthcare is a dynamic system that evolves over time. There are domain specific solutions that can address time series analysis for specific tasks ([91], [147], [55]), but developing a general tool that can find novel multivariate associations between features in time varying data for arbitrary problems is an open challenge. Furthermore, there are many practices that hospitals use and many patient outcomes that describe the overall well-being of the patients (e.g., infection rate, prevalence of chronic lung disease rate, mortality, etc.). To identify possible causations and correlations in the data, one should further account for delays between the time that a hospital adopts a new practice and resulting changes in the outcomes become observable. While there is a lot of literature on data mining in time series data and finding patterns (e.g., [22, 28, 30, 63]), to our knowledge there are no previous methods that simultaneously account for the above mentioned factors.

The second computational approach we employ uses agent-based modeling (ABM) to create models of healthcare practitioners' collaborations. This approach allows researchers to simulate different scenarios and to seek answers for questions that are very hard to answer in real-world situations. The usefulness of an ABM approach depends heavily on the expert knowledge used to construct models that reflect relevant characteristics of the real world. Agent based modeling has been successfully applied in many fields for modeling complex, heterogeneous and distributed systems with many interactions among the entities. Some applications of ABMs can be found in healthcare domain for problems such as supporting expert's decision-making or the coordination of the execution of healthcare activities (e.g., [143], [1]).

This dissertation uses ABMs that build off the work in [27]. In [27] the authors use ABMs to explore potential advantages of QICs over RCTs under various conditions, where agents represent different healthcare institutions. To model healthcare improvement in complex environments with non-linear interactions between healthcare practices, the authors use the metaphor of search on fitness landscapes, popular in the evolutionary computation literature [145]. In this context, healthcare is represented as a clinical fitness landscape where each hospital's (i.,e. agent's) location on the landscape is identified by binary practices used in that hospital (i.e., features) and the height of the landscape at its location in the landscape indicates the patient survival rate at that hospital (i.e., agent's fitness). The differences between hospitals are expressed by the Hamming distances separating them in feature space. As agents seek practices that result in better outcomes they try to move uphill in these landscapes. In [27] the authors use a parametric interaction model ([80, 101–103]) for generating clinical fitness landscapes with randomly generated coefficients on main effects and on different numbers of 2-feature interactions. As the number of interactions increases the landscapes become more rugged and presumably more difficult to search.

In [27] the authors show that search strategies modeled after QICs result in better patient outcomes in nearly all possible scenarios compared to more traditional RCTs, due to a combination of reduced sensitivity to sample size and the ability for QIC agents to respond differently in different local contexts. While their studies showed empirically computational evidence that QICs are better than RCTs, they only used randomly formed teams of fixed sizes for QICs and didn't study in detail what features of QIC team formation result in better outcomes. This topic is further explored in this dissertation. Interestingly, they found that the advantage of QICs vs RCTs increases with an increasing number of interactions between practices. As the number of interactions increases in the landscape, the ranges of fitness values in interaction models change, therefore it is impossible to compare these landscapes based on raw fitnesses. In [27] authors thus use a logistic function to constrain fitnesses to the open interval (0,1). However, this causes many suboptimal peaks to have nearly identical fitnesses to the global optimum, which remained unknown.

One of the challenges in modeling complex problems like improvement in healthcare landscapes is the need for good benchmark problems for testing search algorithms used by ABMs. NK landscapes [62] have been the classic benchmarks for generating fitness landscapes with epistatic interactions and tunable ruggedness. Both the size of the landscape and the number of its local "hills and valleys" can be varied using two parameters: the number of binary features *N* and the maximum degree of epistatic interactions among the features $K + 1$ [60]. *NK* landscapes have been used in many applications (e.g., [2, 36, 82, 112, 129]) and widely studied in theory (e.g., [15, 53, 69, 92, 144]), as they can generate landscapes with tunable ruggedness by varying K. The early versions of NK landscapes only considered the smoothest (K = 0) and the most rugged (N = K) landscapes [61]. There are many versions of NK models, but the classic NK model was first published in [62]. According to [62], in an NK model the fitness is measured as the sum of contributions from each individual feature or "state", where the contribution of each feature depends on K other features from N possible features. In classic NK models K is fixed for all features, although in a generalized model it can vary (e.g., [68]). There are also restricted NK models such as spatially embedded NK models where each feature's contribution is restricted to be only a function of its immediate K spatial neighbors. The problem of finding the location and value of the global optimum of unrestricted *NK* landscapes with $K > 1$ is NP-complete [144] (although for restricted classes one can use dynamic programming [144] [35] or approximation algorithms [144]). While NK landscapes have been widely used as benchmark problems, they have some major limitations. For example, one can not normalize fitness values due to unknown global optima, therefore it is inappropriate to compare results across different landscapes. Another limitation of NK landscapes as benchmarks is that as K increases the ruggedness of the landscape increases in large discrete jumps and it is impossible to incorporate individual interactions among features at a finer level (e.g., see [15]), as is necessary for modeling interactions in healthcare landscapes. Furthermore, NK landscapes have only been defined for binary alphabets. Walsh polynomials ([31, 59]) (defined in detail later in Chapter 4) overcome some of these limitations. In particular, they provide a means for generating landscapes with more smoothly tunable ruggedness. However, they are also only defined for binary alphabets and have unknown global optima. Both NK landscapes and Walsh polynomials are shown to be subsets of general parametric interaction models [15, 59]. Parametric interaction models are easy to define on both discrete and real-valued alphabets, and the interactions are transparent and easy to interpret (unlike in *NK* landscapes and Walsh polynomials). However, finding the global optimum remains an NP-complete problem.

1.2 Outline of This Dissertation

This dissertation is organized as follows: In Chapter 2, we discuss real time series hospital data, including data on social interactions aimed at improving healthcare. We introduce a new method for exploratory analysis of large data sets with time-varying features, where the aim is to automatically discover novel relationships between features (over some time period) that are predictive of any of a number of time-varying outcomes (over some other time period). In chapter 3, we build off the ABM of QICs introduced in [27] to study how various aspects of team formation (e.g., team sizes, how often teams are reformed, amount of data shared among teams, etc.) affect the efficacy of learning. We show (among other things) that teams comprising similar individuals outperform those with more diverse individuals under nearly all conditions, and that this advantage increases with the complexity of the landscape and the level of noise in assessing performance. Examination of data from a network of real hospitals provides encouraging evidence of a high degree of similarity in clinical practices, especially within teams of hospitals engaging in QIC teams. However, our model also suggests that groups of similar hospitals could benefit from larger teams and more open sharing of details on clinical outcomes than is currently the norm. Thus, we propose a secure virtual collaboration system that would allow hospitals to efficiently identify potentially better practices in use at other institutions similar to theirs, without any institutions having to sacrifice the privacy of their own data. In chapter 4, we introduce a new class of benchmarks called *NM* landscapes, where M refers to the Maximum order of epistatic interactions between N features. Like Walsh polynomials, *NM* landscapes are much more smoothly tunable in ruggedness than *NK* landscapes. For all NM landscapes the location and the value of the global maximum is trivially known. For a subset of *NM* landscapes the location and the value of the global minimum is also known, enabling proper normalization of fitnesses. NM landscapes use a natural and transparent representation of epistasis and work with alphabets of any arity, from binary to real-valued. Thus they are well-suited for modeling clinical fitness landscapes. In Chapter 5, we summarize conclusion and propose future work.

Chapter 2

Automated Discovery of Multivariate Associations in Large Time-Varying data sets: a Healthcare Network Application

Manukyan, N., Eppstein, M.J., Horbar, J.D., Leahy, K.A., Kenny, M.J., Mukherjee, S., and Rizzo, D.M. “Exploratory Analysis in Time-Varying data sets: a Healthcare Network Application”, International Journal of Advanced Computer Science, 3(7), 2013.

2.1 Abstract

We introduce a new method for exploratory analysis of large data sets with time-varying features, where the aim is to automatically discover novel relationships between features (over some time period) that are predictive of any of a number of time-varying outcomes (over some other time period). Using a genetic algorithm, we co-evolve (i) a subset of predictive features, (ii) which attribute will be predicted, (iii) the time period over which to assess the predictive features, and (iv) the time period over which to assess the predicted attribute. After validating the method on 15 synthetic test problems, we used the approach for exploratory analysis of a large healthcare network data set. We discovered a strong association, with 100% sensitivity, between hospital participation in multi-institutional quality improvement collaboratives during or before 2002, and changes in the risk-adjusted rates of mortality and morbidity observed after a 1-2 year lag. The results provide indirect evidence that these quality improvement collaboratives may have had the desired effect of improving health care practices at participating hospitals. The proposed approach is a potentially powerful and general tool for exploratory analysis of a wide range of time-series data sets.

2.2 Introduction

The rapid growth of technology has facilitated widespread collection and storage of vast amounts of time-varying data (e.g. [6]). This data undoubtedly contains a wealth of potentially valuable information regarding relationships between various time-varying features and outcomes. However, the very size of these databases is an impediment to knowledge discovery, creating a need for automated exploratory analysis tools. Over recent decades the scientific community has expressed an increasing interest in knowledge discovery in large databases [29], and some exciting progress has been made in this area. For example, a new method for automated discovery of non-parametric associations between pairs of variables was recently proposed and was shown to discover a wide range of functional and non-functional associations [105]. However, it would be computationally prohibitive to extend this method for discovering multivariate associations.

In general, large data sets include many features, only a few of which may interact, potentially in very nonlinear ways, resulting in some association with other outcome features in the data. Thus, identifying the relevant features is a critical aspect of knowledge discovery in large data sets. Evolutionary algorithms provide a particularly attractive approach for feature selection, because they require no pre-determination of the number of features in the optimal feature subset. Genetic algorithms (GAs), in particular, have been widely and successfully applied for feature selection in a variety of problems (e.g., [19], [70], [86], [100], [93]).

However, identifying the correct set of features is only part of the challenge in exploratory data analysis. For example, one may also need to identify which outcome(s) those features are associated with. Indeed, many distinct complex relationships between different feature subsets and different predicted outcomes may be present in the same data set, waiting to be discovered. The problem is compounded with time-series data sets, where there may be time-dependent aspects to the association. There are domain specific solutions that can address this problem for specific tasks ([91], [147], [55]), but developing a general tool that can find novel multivariate associations between features in time varying data for arbitrary problems is a much bigger challenge.

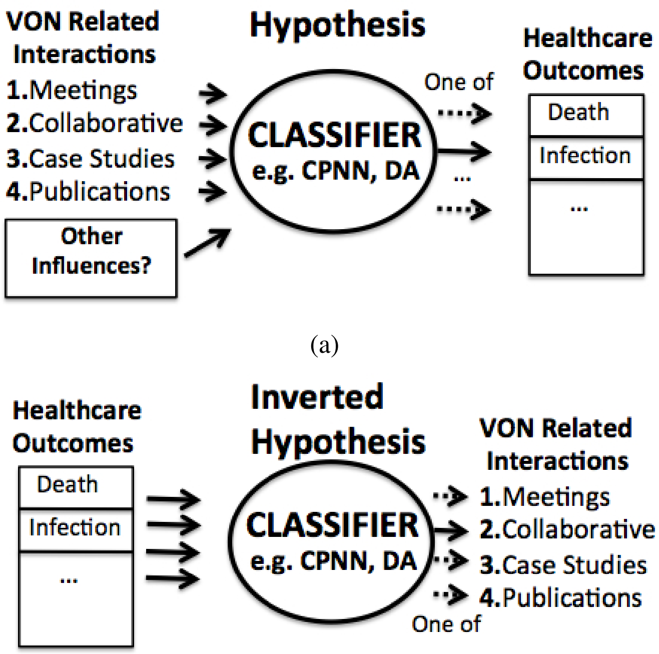
Our motivation in addressing this problem stems from a particular application in the healthcare domain. The Vermont Oxford Network (VON) is a non-profit corporation dedicated to the mission of improving the quality and safety of medical care for newborn infants and their families through a coordinated program of research, education, and networking of neonatal intensive care units (NICUs) at hospitals around the world. Since its inception in 1990, the VON has maintained databases with detailed information about hospital characteristics, treatments, and outcomes for all of the very low birth weight (VLBW) infants (birth weight under 1500 grams) treated at member hospitals around the world (e.g. [45], [44], [11], [46], [43], [108], [109], [83], [148]). These data are used to quantify treatment practices and risk-adjusted morbidity & mortality for VLBW infants treated at NICUs in the VON. While they account for only one percent of births, VLBW infants account for half of infant deaths in the US each year [76]. A major and consistent finding of previous VON database analysis is the dramatic variation in outcomes among NICUs, even after adjusting for differences in case mix among units [44], [46], [43], [108], [109], [83], [148]. Differences in hospitals and unit characteristics such as teaching status, volume or NICU level also fail to explain the large discrepancies in health outcomes [108]. We hypothesize that differences in VON-sponsored activities designed to improve healthcare practices may account for some of these unexplained discrepancies in patient outcomes in VON member hospitals. Of particular interest are VON-sponsored team quality improvement collaboratives, in which interdisciplinary teams from multiple institutions work together to identify, test, implement, and report on innovative evidence-based treatment strategies [51], [50], [107], [48], [47], [98], [88]. In order to explore this hypothesis, we have assembled a large database of VON-sponsored interactions among member hospitals between 1995 and 2010. We seek to discover novel multivariate associations between time-varying VON-sponsored hospital interactions and patient outcomes. Discovering such relationships, if they exist, could potentially have widespread application to managing collaborative healthcare networks, such as the VON, that seek to innovate and spread quality improvement practices between hospitals around the world.

In this paper we propose a genetic algorithm for co-evolving four important aspects of exploratory multivariate time-series analysis: (i) a subset of features to be used as input into some sort of statistical predictor (such as a classifier or regression analysis), (ii) which attribute we can best predict from these features, (iii) a dividing year that partitions the time-series, and (iv) a time lag to be added to the dividing year. Fitness is determined by seeing how well the values of the selected features before the dividing year can be used to predict changes in the selected attribute after the dividing year + lag. In this proof-of-concept study, we first validate the approach using synthetic data, and then apply the method to a subset of the VON data.

2.3 Methods

We propose a new method that uses a Genetic Algorithm (GA) to co-evolve the inputs and output to a fitness function based on a statistical predictor, seeking causal associations in large time-varying data sets with multiple input features and potential prediction attributes. In this paper we focus on classification predictors, although one could easily employ other types of predictors (such as multiple regression). For brevity, we refer to this method as GAMET (Genetic Algorithm for Multivariate Exploration of Time-varying data).

In the general problem, the hypothesis is that there is some sort of causal relationship between a set of features that affect the value of some outcome attribute over some time period in the future. For example, we hypothesize that interactions between hospitals in the Vermont Oxford Network (e.g., as evidenced by participation in multi-institutional team quality improvement collaboratives, co-authored publications, case study presentations, and attendance at annual meetings) can influence future health outcomes at these hospitals (e.g., probability of patient death, infection, or other morbidity). However, even assuming this causal influence is true, there are doubtless a number of other (non-VON related) influences that affect the healthcare outcomes at these hospitals (see Fig. 2.1, top). Thus, it is not realistic to expect that we will be able to predict healthcare outcomes based on knowledge of the VON interactions alone. Furthermore, the number of hospitals that actively participate in the more intense types of VON interactions (such as team collaboratives and co-authorship on scientific studies) is much smaller than the number of member hospitals that don't actively participate, so these classes are very imbalanced. Consequently, for this application we seek to do the prediction in the opposite direction (see Fig. 2.1, bottom). That is, given a knowledge of time-varying healthcare outcomes at various hospitals, can we predict which hospitals actively participated in VON-sponsored interactions (even if we cannot determine which hospitals did *not* actively participate)?



(a)

(b)

Figure 2.1: a) Hypothesis of causality. b) Inverted hypothesis tested by the classifier.

In a problem like this, where we hope to infer causal relationships, it is important to take the time-varying nature of the data into account. For example, if a hospital participates in a team collaborative designed to reduce infection rates, then one would hope to see infection rates decrease at that hospital at some time in the future, although there may be a time lag between when the collaborative activity took place and when measurable changes in

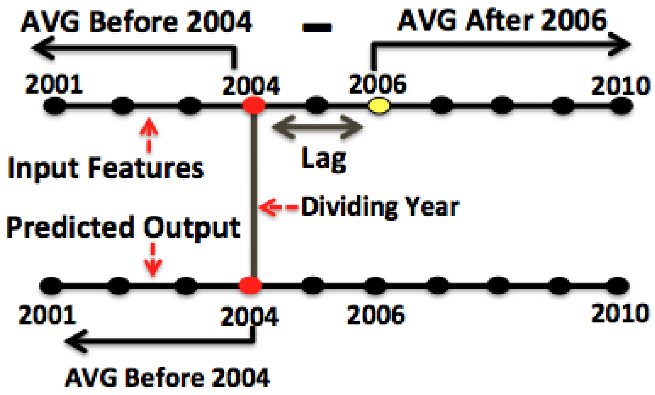


Figure 2.2: Information is extracted and aggregated from the time-series data relative to a dividing year (2004, in this example) and lag (2 years, in this example). Specifically, we compute the change in average input feature values after the dividing year+lag, relative to during or before the dividing year. These are used to try to classify the values of the predicted output, averaged over all years during or prior to the dividing year. Here, the terms "input" and "output" are relative to the classifier used in the inverted hypothesis (see Fig. 2.1).

infection rate can be detected. We handle this time component by looking at the change in health outcomes, averaged before and after a given points in time, relative to some "dividing year" and possibly with an intervening time lag, and see if we can use this to predict the presumed causal attribute (level of participation in VON-sponsored activities) before the dividing year (as illustrated in Fig. 2.2 for a dividing year of 2004 and a time lag of 2 years).

Thus, we desire to co-estimate three types of information simultaneously: which features to use as input to the classifier, what dividing year and lag to use in processing the time-series data, and which attribute to try to predict. The binary chromosomes used in GA-MET thus include genes associated with each of these three parts (see Fig. 2.3). For feature selection, we are using binary flags that indicate whether the given feature is included in the final features subset or not. To evolve the time series component we evolve the dividing year and lag, both of which are represented as gray-coded integers in the chromosome. Finally, a gray-coded “participation index” specifying which single attribute (from a list of potentially predicted attributes) is to be predicted.

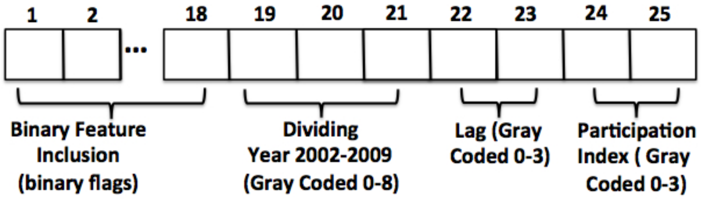

1
2
...
18
19
20
21
22
23
24
25
Binary Feature Inclusion (binary flags)
Dividing Year 2002-2009 (Gray Coded 0-8)
Lag (Gray Coded 0-3)
Participation Index (Gray Coded 0-3)

Figure 2.3: Example GAMET chromosome for the VON data, allowing for up to 18 possible features, dividing year $\in\{2002,...,2009\}$, lag$\in\{0,...,3\}$ years, and one of four possible attributes to predict (specified by a participation index).

To calculate the fitness of an individual, we first process the data for the included features, using the dividing year and lag as described above (labeled as time series extraction and aggregation in Fig. 2.4). We then pass these time-processed features as inputs to the classifier, and compare the predicted classes to class outcomes of the attribute specified by the participation index, averaged prior to the dividing year. The data is divided into training and testing sets, using a parameter to control the percentage of the data used for training (80% for our experiments). We use Latin hypercube sampling to ensure adequate distribution of samples in the training and testing sets for this highly unbalanced classification problem. After the training phase we evaluate the classifier performance using the confusion matrix, which shows the number of correctly and incorrectly classified samples in each class (see Fig. 2.4).

For our VON data set we are using two classes for all predicted outputs: a “positive” (P) classification means that we are predicting that a particular hospital participated in the specified activity before the dividing year, whereas a “negative” (N) classification means we are predicting the hospital didn't participate in the specified activity. The fitness is calculated using the following formula:

$$fitness = \frac{FP}{(FP+TN)} + \frac{FN}{(FN+TP)} + \frac{(FP+FN)}{2(TP+FP+TN+TP)} \tag{2.1}$$

where FP is the number of false positives, TP is the number of true positives, FN is the number of false negatives and TN is the number of true negatives. The first two terms represent the proportion of samples in each class that were classified incorrectly, whereas the last term is the proportion of the overall misclassified samples. This fitness function thus takes into consideration both the overall prediction rate and the individual class prediction rates (the latter is helpful for unbalanced classes). We would like to note that there is some stochasticity involved in the calculation of the fitness function (due to the Latin hypercube sampling and any stochasticity possibly associated with classifier), which can result in slightly different fitness values being evaluated for the same chromosome on different occasions.

We employ two different classifiers in this paper. For the synthetically generated data set, we were able to use a naïve Bayes quadratic discriminant analysis (DA) classifier. However, because the VON data set violated so many assumptions of the DA, for this application we used a non-parameter counterpropagation artificial neural network (CPNN) classifier [40]. The overall architecture of the approach, illustrated for the VON data set, is shown in Fig. 2.4, where the co-evolved entities are indicated in red.

Table 2.1: GA parameters used in this study.

Parameter	**Value**
Population Type	bitstring
Population Size	500
Generations	100
Crossover Fcn	scattered
Mutation Fcn	{uniform, p = 0.04}
Crossover Fraction	0.8
Elite Count	1
Selection Fcn	{tournament, size = 4}

Table 2.2: CPNN parameters used in this study.

Parameter	**Value**
Learning rate	0.7
Bias	0.1
Mean Square Error to stop training	0.001

2.4 Experiments

2.4.1 Synthetic data

In order to test the capability of GAMET for co-evolving correct feature sets of varying sizes, attribute to predict, year, and lag, we created synthetic data sets for 15 test problems, as follows.

We first generated 5 random "true" combinations of dividing year (selected uniformly from 2002..2009), lag (selected uniformly from 0..3 years), and index for the attribute to predict (selected uniformly from 0..3). We next generated 15 random multivariate expression trees in 3 sets of varying levels of difficulty; 5 expressions contained 2 variables, 5

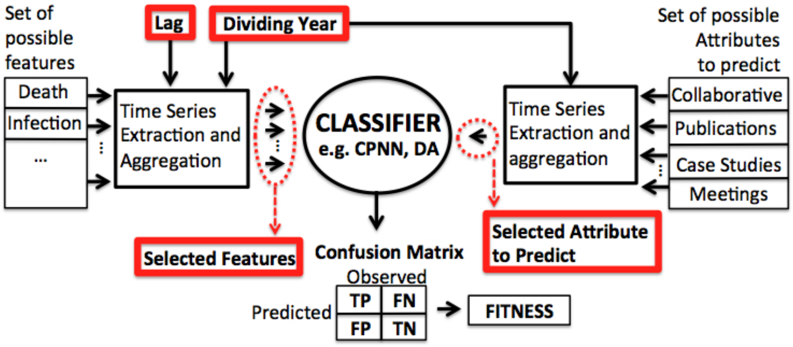


Figure 2.4: Overall architecture of the approach, illustrated for use with the VON data set. Items outlined in red are co-evolved by GAMET.

contained 3 variables, and 5 contained 8 variables. For each of these 15 test problems, we generated a 300×100 matrix of uniformly distributed random real numbers in the range (0,1), representing synthetic data for 300 cases, each with 100 feature variables (e.g., synthetic values for 100 heath outcomes at 300 hospitals). The expression trees were generated using a function set of $\{+, -, *, exp, <, >, ==\}$ and were constructed so as to return binary class outcomes such that at most 2/3 of the outcomes had the same value. The expression trees were generated using a terminal set comprising 100 distinct real-valued variables (corresponding to the 100 feature columns in the synthetic data sets), as well as integer constants {1,2,3}. Each set of 5 expression trees with the same number of variables was associated with the set of 5 combinations of year, lag, and index of the attribute to be predicted, created as described above. The resulting specifications for these 15 test problems are outlined in Table 2.4, column 2.

For each of the 15 random problems, we then created a synthetic 300×128 outcomes matrix, where the 128 columns in this matrix correspond to all combinations of 4 possible attributes to predict (e.g., synthetic values for participation in 4 types of VON-related interactions), 8 possible dividing years, and 4 possible lags. All 96 columns in the outcomes matrix corresponding to the 3 incorrect attributes to predict (for all 8 dividing years and all 4 possible lags) were initialized to uniform random binary class outcomes. However, the remaining 32 columns associated with the correct attribute to predict (for all 8 dividing years and all 4 lags), were initialized to the "true" predicted binary class outcomes associated with the 15 random problems. These "true" outcomes were calculated by evaluating the expression trees using the columns from the synthetic data matrix corresponding to the feature variables in the expression trees. Lastly, we added noise to 31 of these 32 columns, proportional to the Hamming distance (H) between the 5-bit gray-coded sequences representing their dividing years (3-bits) and lags (2-bits) and the 5-bit vector representing the "true" dividing year and lag. Specifically, we overwrote $30 \times H$ bits in each of these columns with random binary values.

This algorithm thus creates a synthetic data set that has known associations between a subset of feature vectors and one of the attributes to predict. By design this relationship has a perfect association when the dividing year and the lag exactly match the "true" target values, but the level of added random noise increases as the dividing year and lag get farther from the target values, as one might expect to see in real time series data.

2.4.2 VON data set

VON-related interactions

We assimilated a large database of VON-facilitated interactions between hospitals for the years 1995 through 2010. During this time, the VON network grew from around 100 hospitals to 850 hospitals. Here, we report on four specific types of VON-sponsored interactions: (i) participation in VON annual meetings; (ii) preparation of case studies that were presented at VON meetings; (iii) participation in VON-sponsored team collaboratives, which are 2-year long team projects where multidisciplinary quality improvement teams from participating hospitals work together to identify and implement potentially better health practices, and (iv) co-authorship on publications resulting from VON-related activities. It should be noted that the level of participation in these four types of interactions is quite variable, with many member hospitals not actively participating in any of these types of VON-sponsored interactions. On average, in any given year only {53.2%, 11.5%, 8.3%, and 21.5%} of all VON member hospitals participated in these four types of activities, respectively. Thus, although we have quantitative information on the amount of participation in each of these activities, for this preliminary study we have binarized the annual participation in these four types of interactions for each member hospital. Our initial goal is to see if changes in health outcomes are associated with any level of participation, in any of these types of VON-facilitated interactions. I.e., these four types of VON-sponsored interactions comprise four potential "attributes to predict", where the predicted values are the binary classes representing participation or non-participation. After the creation of this database, all identifying information was removed, to ensure member hospital privacy.

VON health outcomes

The VON maintains an extensive database of over 200 types of annual health outcomes at all member hospitals. In this preliminary study, we are focusing on only 18 risk-adjusted measures (see table 2.3) over the period 2001 through 2010, representing the health outcomes of over half a million VLBW infants. The risk adjusted outcome measures are recorded as observed divided by expected values of the outcome, where expected values vary with the number of patients at the hospital. These particular features were identified by

VON staff as ones they thought had strong potential to have been impacted by VON-related interactions, based on collaborative studies they had sponsored during this time period. I.e., we want to see if subsets of these 18 real-valued features can be used to classify individual hospitals as participants or non-participants in any of the 4 types of VON-sponsored interactions described in Section 2.4.2. The distribution of health outcomes in the real VON data violates assumptions of normality and independence. Preliminary testing, using the real VON health outcome features described in Section 2.4.2 with synthetically generated known associations to class outcomes, confirmed that the parametric DA classifier was not able to correctly classify known outcomes associated with these data, whereas the non-parametric CPNN was. Thus, as mentioned previously, we used CPNN-based fitness in the co-evolutionary method applied to the VON data. All hospital data was provided to us in a completely anonymized manner, to ensure member hospital privacy.

2.4.3 Experimental design

For the 15 synthetic problems described in Section 2.4.1, we ran 10 replicates of the GA, using the DA-based fitness function. For the actual VON data described in Section 2.4.2, we ran 10 replicates of the GA, using the CPNN-based fitness function. Because both the DA and the CPNN can still classify well even with a certain number of excess features given as inputs, we subsequently intersected the feature sets of the best individuals resulting from each of the 10 replicates. The results of these experiments are described in the following section.

Table 2.3: Health outcomes used as possible features in our analysis of the VON data.

#	**Description**
1	Any Late Infection
2	Chronic Lung Disease
3	Chronic Lung Disease before 33 Weeks
4	Coagulase Negative Staph
5	Mortality
6	Mortality or Morbidity
7	Fungal Infection
8	Intraventricular Hemorrhage
9	Mortality Excluding Early Deaths
10	Bacterial Pathogen after Day 3
11	Necrotizing Enterocolitis
12	Necrotizing Enterocolitis, where occurred
13	Nosocomial Infection
14	Pneumothorax
15	Cystic Periventricular Leukomalacia
16	Retinopathy of Prematurity
17	Severe Intraventricular Hemorrhage
18	Severe Retinopathy of Prematurity

2.5 Results

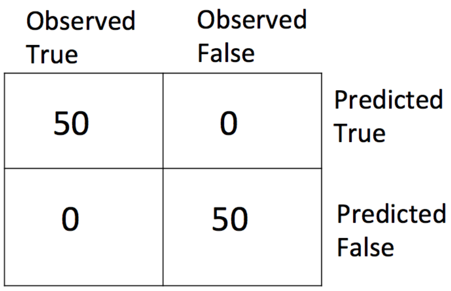
In all 10 replications of each of the fifteen 100-feature synthetic problems GAMET was able to correctly identify the dividing year, lag, which attribute to predict (labeled ”output”), and all of the 2, 3, or 8 true features (see table 2.4, compare columns 2 and 3), using the DA-based fitness function.

Table 2.4: Experimental results on the 15 synthetic test problems. The forth column shows the means and standard deviations of the number of features found by GAMET for the best individuals from each of 10 runs, and the fifth column shows the number of excess features in the intersections of these feature subsets from the 10 runs.

#	**True** year, lag, output, #features	**Found** year, lag, output, #true feat.	**#Feat.** mean± std	**Excess** #features found in ∩
1	2002, 2, 3, 2	2002, 2, 3, 2	46±4	0
2	2003, 1, 1, 2	2003, 1, 1, 2	45±5	0
3	2007, 0, 2, 2	2007, 0, 2, 2	46±6	0
4	2005, 2, 4, 2	2005, 2, 4, 2	44±4	0
5	2004, 1, 3, 2	2004, 1, 3, 2	45±5	1
6	2002, 2, 3, 3	2002, 2, 3, 3	48±5	0
7	2003, 1, 1, 3	2003, 1, 1, 3	47±6	1
8	2007, 0, 2, 3	2007, 0, 2, 3	47±6	1
9	2005, 2, 4, 3	2005, 2, 4, 3	48±6	1
10	2004, 1, 3, 3	2004, 1, 3, 3	47±5	1
11	2002, 2, 3, 8	2002, 2, 3, 8	49±7	3
12	2003, 1, 1, 8	2003, 1, 1, 8	48±6	1
13	2007, 0, 2, 8	2007, 0, 2, 8	52±8	3
14	2005, 2, 4, 8	2005, 2, 4, 8	50±7	2
15	2004, 1, 3, 8	2004, 1, 3, 8	51±8	1

As the number of true features increased, the tendency of GAMET to return excess features also increased (table 2.4, column 4), since the DA can accommodate excess features

Table 2.5: Confusion matrix for the intersection of 10 best individuals for all 15 tests.

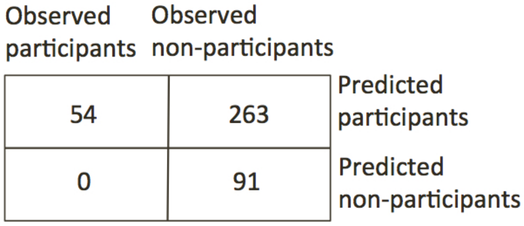


Observed True	Observed False	
50	0	Predicted True
0	50	Predicted False

(but simply not give them much weight). However, the intersections of the feature sets in the 10 replications contained relatively few excess features (table 2.4, column 5). When run with features in the intersection of the 10 runs, all the 100 testing data points were correctly classified (table 2.5). These results demonstrate that the system is able to co-evolve feature subsets that include all the correct features (intersecting several of these sets removes excess features), the correct attribute to classify, the correct dividing year, and the correct lag in synthetic time-series data with known relationships between input features and attribute to classify.

On the VON data set, all 10 runs consistently returned a dividing year of 2002, and discovered that participation in VON-sponsored team collaboratives was the attribute that could most accurately be classified. In 7 of the 10 runs, the lag was determined to be 2 years, whereas in the remaining 3 runs the lag was determined to be 1 years. The health outcome features selected as input to the CPNN-based fitness function were also relatively consistent between the 10 runs (see Fig. 2.5). However, since the CPNN can do robust predictions even when given a few excess inputs, we then searched for consensus in the selected features between the different runs.

Table 2.6: Confusion matrix for the best individual found by GAMET on the VON data. Here, “participants” refers those hospitals who participated in VON-sponsored quality improvement collaboratives during or before to 2002.



Observed participants	Observed non-participants	
54	263	Predicted participants
0	91	Predicted non-participants

In all cases, the CPNN was able to predict the “true positives” in the smaller class (participants) with 100% accuracy (i.e., based on the selected health outcomes, the CPNN could correctly predict which hospitals *had* participated in a VON-sponsored team collaborative during or before the dividing year) (see table 2.6, column 1). However, the classifier was not able to use the selected health outcomes to accurately predict the “true negatives” (hospitals that didn’t participate in any VON-sponsored team collaboratives during or before the dividing year) (see table 2.6, column 2). In other words, the identified classifier has high sensitivity, but low specificity. We assessed the overall classification accuracy in prediction participation in a VON-sponsored team collaborative, using health outcome feature sets that included the top $n \in \{4, 8, 9, 11, 12, 13, 16\}$ most consistently selected features, based on the consensus features selected in $\{100\%, 90\%, 80\%, 70\%, 30\%, 20\%, 10\%\}$ of the replicates, respectively (i.e., those features whose frequency bars are at or above the horizontal dotted lines in Fig. 2.5), using a dividing year of 2002 and a lag of 2 years. We report the resulting percentage accuracies to the right of Fig. 2.5. Four features (5, 6, 8, and 14) occurred in the selected features of all 10 replicates, but the highest prediction accuracy (33%) was obtained when using the 9 features (2, 5, 6, 7, 8, 9, 10, 11, 12, 14, and 17) that were found in at least 7 of the 10 replicates; this 9-feature set also coincides with the best single individual found in the 10 runs (Fig. 2.5, red asterisks). The confusion matrix for this individual is shown in Table 2.6. Note that differences in these percent accuracies only reflect the differences in the specificity of the classifier, since all had perfect sensitivity. Conversely, we also found that the overall classification accuracy dropped dramatically to only 16%-18% when predicting from any 3 of the top 4 features, indicating that all four of these are important predictive features.

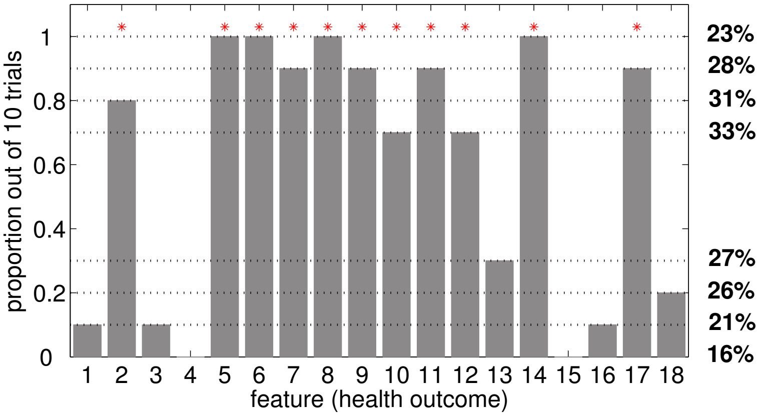


Figure 2.5: Experimental results on the VON data set. The bars indicate the frequency with which each of the individual features was selected in 10 GAMET trials. The red asterisks near the top indicate the features selected in the single best individual. The percentages to the right indicate overall classification accuracy of the CPNN from the consensus features with frequencies at or above each of the horizontal dotted lines. Feature numbers correspond to those shown in Table 2.3.

2.6 Discussion and Conclusions

In this paper, we introduce a method for exploratory analysis of large data sets with time-varying features. Such data sets may contain information about many different potential relationships between features and outcomes. The aim is to automatically discover novel relationships between features (over some time period) that are predictive of any of a number of time-varying outcomes (over a different time period), but where the specific features, outcomes, and time periods are not known in advance. The application that motivated this study concerns exploratory analysis of a large healthcare network data set, comprising various types of time-varying interactions between subsets of hospitals in the Vermont Oxford Network (VON) and a variety of annual health outcomes at those hospitals.

The approach we take uses a Genetic Algorithm for Multivariate Exploration of Time-varying data (GAMET), in which we co-evolve (i) a subset of health outcomes, (ii) one of four types of VON-sponsored interactions to consider, (iii) the maximum "dividing" year up to which we consider these VON-sponsored interactions, and (iv) how many years time lag after the dividing year before which we assess changes in the health outcomes.

We first validated that GAMET was able to select the correct features, outcomes, dividing year, and lag in 15 synthetically designed problems with 2, 3, and 8 non-linearly interacting features with known associations to a specific binary-valued attribute. For these synthetic problems we assessed fitness based on the classification accuracy of a naïve Bayes quadratic discriminant analysis classifier.

We then conducted preliminary exploration of the actual VON data set with 18 potential health outcome features, 4 types of VON-sponsored interactions, 8 possible dividing years, and 4 possible lags, representing a search space of over 33 million possible combinations of solutions. Due to the non-parametric nature of this actual data set, we assessed fitness based on the classification accuracy of non-parametric counter-propagation artificial neural network classifier. In addition, because the participation classes were highly unbalanced, we used Latin hypercube sampling to determine how to subdivide the data into appropriate training and testing sets.

The strongest association so far discovered by GAMET in the VON data set was between participation in VON-sponsored team quality improvement collaboratives during or before 2002, and changes in the risk-adjusted rates of mortality and morbidities including intraventricular hemorrhage and pneumothorax (collapsed lung) that were observed after 2003 or 2004, relative to these rates during or before 2002. Using changes in only 4 health outcomes selected by GAMET, we achieved 100% sensitivity in predicting which hospitals had participated in these collaboratives in 2002 or earlier.

From a clinical standpoint, these results are interesting, because the team collaboratives sponsored by the VON up through 2002 had included teams focussing on intraventricular hemorrhage, chronic lung disease and ventilation, respiratory care and management, and a number of healthcare practices designed to positively impact overall rates of morbidity & mortality. Subsequent individual analysis of three of the four outcomes (risk adjusted rates of mortality, morbidity & mortality, and intraventricular hemorrhage) showed a slight average improvement in NICQ hospitals, while the fourth outcome (risk adjusted rate of pneumothorax) showed slight average degradation in NICQ hospitals; however, there was no statistically significant difference between changes in these outcomes for NICQ vs. non-NICQ hospitals for any of these four outcomes (t-test, $p > 0.79$).

The identified lag of 1-2 years is a reasonable amount of time one would expect such changes in health practices to be implemented, and the health impacts of these changes observed, in the annually-updated health outcome records.

Our results on the VON data had relatively low specificity, however. The best individual returned by GAMET was still only able to achieve an overall classification accuracy of 33%, because the classifier was not able to accurately predict which hospitals had *not* participated in VON-sponsored interactions during or before 2002, based on the changes in health outcomes after 2003 or 2004. This result is actually to be expected, because there are many changes in healthcare practices at VON member hospitals that were independent of participation in VON-sponsored activities (and are consequently not in our database) that are expected to contribute to changes in health outcomes.

Having established proof-of-concept for the method, we now plan to apply GAMET to a more complete set of health outcome features and VON-sponsored interactions aimed at stimulating improvements in healthcare practices. We will then more closely examine the specific nature of the relationships embedded in the associations discovered by GAMET. For example, we intend to use genetic programming (GP) for symbolic regression, using GAMET-selected features as variables in the GP terminal set (much as in [19]).

We can also envision many ways in which to improve the GAMET algorithm itself. For example, since the two types of classifiers employed here (the DA and the CPNN) can be trained to ignore excess features, the features selected by GAMET also contained excess features. Consequently, we applied a post-processing step to further reduce the final feature sets, by looking for features common to the selected feature sets from different GAMET replicates. Others have reported promising results in GA-based feature selection by actually embedding set intersection directly into the crossover operator [66], [19]. Although we found that strict set intersection was too aggressive in reducing features in the VON application, we plan to explore whether a probabilistic application of a “softer” form of multi-set intersection (i.e., including all elements that occur in a certain percentage of parents) in multi-parent crossover could help improve feature selection in GAMET, and therefore preclude the need for the post-processing of multiple replicates, as done here. In addition, the current version of GAMET only allows for the evolution of a single dividing year. We plan to explore whether it may prove more powerful to apply the evolved lag directly to the hospital-specific years of participation for selected types of VON interactions.

Although the proposed method was originally developed for analysis of the VON healthcare network data set described here, the GAMET approach is a potentially powerful and general tool for exploratory analysis of a wide range of time-series data sets. Future work will include the application of GAMET to time-vary problems in a variety of other domains (such as those in [6]).

2.7 Acknowledgments

This work was funded in part by the NIH Eunice Kennedy Shriver National Institute of Child Health & Human Development award 1R21HD068296. We thank the staff of the Vermont Oxford Network for their help in assembling the data sets.

Bibliography

[1] Roberta Annicchiarico, Ulises Cortés, and Cristina Urdiales. *Agent Technology and E-health*. Springer, 2008.

[2] Ignacio Arnaldo, Iván Contreras, David Millán-Ruiz, J Ignacio Hidalgo, and Natalio Krasnogor. Matching island topologies to problem structure in parallel evolutionary algorithms. *Soft Computing*, pages 1–17, 2013.

[3] Lea R Ayers, Suzanne C Beyea, Marjorie M Godfrey, Doreen C Harper, Eugene C Nelson, and Paul B Batalden. Quality improvement learning collaboratives. *Quality Management in Healthcare*, 14(4):234–247, 2005.

[4] Rosa R Baier, David R Gifford, Gail Patry, Sara M Banks, Therese Rochon, Debra DeSilva, and Joan M Teno. Ameliorating pain in nursing homes: a collaborative quality-improvement project. *Journal of the American Geriatrics Society*, 52(12):1988–1995, 2004.

[5] David W Baker, Steven M Asch, Joan W Keesey, Julie A Brown, Kitty S Chan, Geoffrey Joyce, and Emmett B Keeler. Differences in education, knowledge, self-management activities, and health outcomes for patients with heart failure cared for under the chronic disease model: the improving chronic illness care evaluation. *Journal of cardiac failure*, 11(6):405–413, 2005.

[6] A.S. Banks. Cross-national time-series data archive (cnts) 1815-2007. *Databanks International, Jerusalem, Israel*, 2008.

[7] Alberto Barceló, Elizabeth Cafiero, Melanie de Boer, Alejandro Escobar Mesa, Marcelina García Lopez, Rosa Aurora Jiménez, Agustín Lara Esqueda, José Antonio Martinez, Esperanza Medina Holguin, Micheline Meiners, et al. Using collaborative learning to improve diabetes care and outcomes: The vida project. *Primary care diabetes*, 4(3):145–153, 2010.

[8] Paul B Batalden and Frank Davidoff. What is quality improvement and how can it transform healthcare? *Quality and Safety in Health Care*, 16(1):2–3, 2007.

[9] Jordana T Bell, Nicholas J Timpson, N William Rayner, Eleftheria Zeggini, Timothy M Frayling, Andrew T Hattersley, Andrew P Morris, and Mark I McCarthy. Genome-wide association scan allowing for epistasis in type 2 diabetes. *Annals of human genetics*, 75(1):10–19, 2011.

[10] Rob Benedetti, Barb Flock, Steve Pedersen, et al. Improved clinical outcomes for fee-for-service physician practices participating in a diabetes care collaborative. *Joint Commission Journal on Quality and Patient Safety*, 30(4):187–194, 2004.

[11] I.M. Bernstein, J.D. Horbar, G.J. Badger, A. Ohlsson, A. Golan, et al. Morbidity and mortality among very-low-birth-weight neonates with intrauterine growth restriction. *American journal of obstetrics and gynecology*, 182(1):198–206, 2000.

[12] DM Berwick. Broadening the view of evidence-based medicine. *Quality and Safety in Health Care*, 14(5):315–316, 2005.

[13] Sebastian Bonhoeffer, Colombe Chappey, Neil T Parkin, Jeanette M Whitcomb, and Christos J Petropoulos. Evidence for positive epistasis in hiv-1. *Science*, 306(5701):1547–1550, 2004.

[14] Penny Bundy. Using drama in the counselling process: the moving on project. *Research in drama education*, 11(1):7–18, 2006.

[15] Jeffrey Buzas and Jeffrey Dinitz. An analysis of NK landscapes: Interaction structure, statistical properties and expected number of local optima. *IEEE Transactions on Evolutionary Computation*, in press, DOI10.1109/TEVC.2013.2286352, 2014.

[16] Pilar Caamaño, Abraham Prieto, José Antonio Becerra, Francisco Bellas, and Richard J Duro. Real-valued multimodal fitness landscape characterization for evolution. In *Neural Information Processing. Theory and Algorithms*, pages 567–574. Springer, 2010.

[17] P.R. Cohen and H.J. Levesque. Teamwork. *Nous*, pages 487–512, 1991.

[18] A. Dechartres, I. Boutron, L. Trinquart, et al. Single-center trials show larger treatment effects than multicenter trials: Evidence from a meta-epidemiologic study. *Annals of internal medicine*, 155(1):39, 2011.

[19] D. DeHaas, J. Craig, C. Rickert, P. Haake, K. Stor, and M.J. Eppstein. Feature selection and classification in noisy epistatic problems using a hybrid evolutionary approach. *poster and published extended abstract accepted for Genetic and Evolutionary Computation Conference (GECCO)*, 2007.

[20] OM Dekkers, Erik von Elm, Ale Algra, JA Romijn, and JP Vandenbroucke. How to assess the external validity of therapeutic trials: a conceptual approach. *International journal of epidemiology*, 39(1):89–94, 2010.

[21] J. Denrell and C. Liu. Top performers are not the most impressive when extreme performance indicates unreliability. *Proceedings of the National Academy of Sciences*, 109(24):9331–9336, 2012.

[22] Hui Ding, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, and Eamonn Keogh. Querying and mining of time series data: experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment*, 1(2):1542–1552, 2008.

[23] Lori Ebert, Lisa Amaya-Jackson, Jan M Markiewicz, Cassandra Kisiel, and John A Fairbank. Use of the breakthrough series collaborative to support broad and sustained use of evidence-based trauma treatment for children in community practice settings. *Administration and Policy in Mental Health and Mental Health Services Research*, 39(3):187–199, 2012.

[24] Judith A Effken. Different lenses, improved outcomes: a new approach to the analysis and design of healthcare information systems. *International journal of medical informatics*, 65(1):59–74, 2002.

[25] Margaret J Eppstein and Paul Haake. Very large scale relieff for genome-wide association analysis. In *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 112–119, 2008.

[26] M.J. Eppstein and P.D.H. Hines. A random chemistry; algorithm for identifying collections of multiple contingencies that initiate cascading failure. *IEEE Transactions on Power Systems*, 27(3):1698–1705, 2012.

[27] M.J. Eppstein, J.D. Horbar, J.S. Buzas, and S.A. Kauffman. Searching the clinical fitness landscape. *PLoS ONE*, 7(11):e49901, 2012.

[28] Philippe Esling and Carlos Agon. Time-series data mining. *ACM Computing Surveys (CSUR)*, 45(1):12, 2012.

[29] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.

[30] Usama M Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy. Advances in knowledge discovery and data mining. 1996.

[31] Stephanie Forrest and Melanie Mitchell. The performance of genetic algorithms on walsh polynomials: Some anomalous results and their explanation. In *Proceedings of the 4th International Cinference on Genetic Alogarithms*, pages 182–189. San Mateo, CA: Morgan Kaufmann, 1991.

[32] Sarah W Fraser and Trisha Greenhalgh. Coping with complexity: educating for capability. *BMJ*, 323(7316):799–803, 2001.

[33] Mingxin Gan and Rui Jiang. Constructing a user similarity network to remove adverse influence of popular objects for personalized recommendation. *Expert Systems with Applications*, 2013.

[34] Mingxin Gan and Rui Jiang. Improving accuracy and diversity of personalized recommendation through power law adjustments of user similarities. *Decision Support Systems*, 2013.

[35] Yong Gao and Joseph C. Culberson. An analysis of phase transition in NK landscapes. *Journal of Artificial Intelligence Research*, 17(1):309–332, 2002.

[36] Ilaria Giannoccaro. Complex systems methodologies for behavioural research in operations management: NK fitness landscape. In *Behavioral Issues in Operations Management*, pages 23–47. Springer, 2013.

[37] R.J. Gray. A bayesian analysis of institutional effects in a multicenter cancer clinical trial. *Biometrics*, pages 244–253, 1994.

[38] Jon W Gregersen, Kamil R Kranc, Xiayi Ke, Pia Svendsen, Lars S Madsen, Allan Randrup Thomsen, Lon R Cardon, John I Bell, and Lars Fugger. Functional epistasis on a common mhc haplotype associated with multiple sclerosis. *Nature*, 443(7111):574–577, 2006.

[39] R. Guimera, B. Uzzi, J. Spiro, and L.A.N. Amaral. Team assembly mechanisms determine collaboration network structure and team performance. *Science*, 308(5722):697–702, 2005.

[40] R. Hecht-Nielsen. Counterpropagation networks. *Applied optics*, 26(23):4979–4983, 1987.

[41] L. Hong and S.E. Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the United States of America*, 101(46):16385–16389, 2004.

[42] J.D. Horbar. The vermont oxford network: evidence-based quality improvement for neonatology. *Pediatrics*, 103(Supplement):350, 1999.

[43] J.D. Horbar, G.J. Badger, J.H. Carpenter, A.A. Fanaroff, S. Kilpatrick, M. LaCorte, R. Phibbs, R.F. Soll, et al. Trends in mortality and morbidity for very low birth weight infants, 1991–1999. *Pediatrics*, 110(1):143, 2002.

[44] J.D. Horbar, G.J. Badger, E.M. Lewit, J. Rogowski, P.H. Shiono, et al. Hospital and patient characteristics associated with variation in 28-day mortality rates for very low birth weight infants. *Pediatrics*, 99(2):149, 1997.

[45] J.D. Horbar et al. The vermont-oxford neonatal network: integrating research and clinical practice to improve the quality of medical care. In *Seminars in perinatology*, volume 19, page 124, 1995.

[46] J.D. Horbar and J.F. Lucey. Evaluation of neonatal intensive care technologies. *The Future of Children*, pages 139–161, 1995.

[47] J.D. Horbar, P.E. Plsek, and K. Leahy. Nic/q 2000: establishing habits for improvement in neonatal intensive care units. *Pediatrics*, 111(Supplement):e397, 2003.

[48] J.D. Horbar, P.E. Plsek, J.A. Schriefer, and K. Leahy. Evidence-based quality improvement in neonatal and perinatal medicine: the neonatal intensive care quality improvement collaborative experience. *Pediatrics*, 118(Supplement):S57, 2006.

[49] J.D. Horbar, J. Rogowski, P.E. Plsek, P. Delmore, W.H. Edwards, J. Hocker, A.D. Kantak, P. Lewallen, W. Lewis, E. Lewit, et al. Collaborative quality improvement for neonatal intensive care. *Pediatrics*, 107(1):14–22, 2001.

[50] J.D. Horbar, J. Rogowski, P.E. Plsek, P. Delmore, W.H. Edwards, J. Hocker, A.D. Kantak, P. Lewallen, W. Lewis, E. Lewit, et al. Collaborative quality improvement for neonatal intensive care. *Pediatrics*, 107(1):14–22, 2001.

[51] J.D. Horbar, R.F. Soll, and W.H. Edwards. The vermont oxford network: a community of practice. *Clin Perinatol*, 37(1):29–47, 2010.

[52] J.D. Horbar, R.F. Soll, and W.H. Edwards. The Vermont Oxford Network: A community of practice. *Clinics in perinatology*, 37(1):29, 2010.

[53] Wim Hordijk. Correlation analysis of coupled fitness landscapes. In *Recent Advances in the Theory and Application of Fitness Landscapes*, pages 369–393. Springer, 2014.

[54] R.I. Horwitz, B.H. Singer, R.W. Makuch, and C.M. Viscoli. Can treatment that is helpful on average be harmful to some patients? A study of the conflicting information needs of clinical inquiry and drug regulation. *Journal of Clinical Epidemiology*, 49(4):395–400, 1996.

[55] T.K. Jenssen, W. Kuo, T. Stokke, and E. Hovig. Associations between gene expressions in breast cancer and patient survival. *Human genetics*, 111(4):411–420, 2002.

[56] K. Johnell and I. Klarin. The relationship between number of drugs and potential drug-drug interactions in the elderly: A study of over 600000 elderly patients from the swedish prescribed drug register. *Drug Safety*, 30(10):911–918, 2007.

[57] Terry Jones and Stephanie Forrest. Fitness distance correlation as a measure of problem difficulty for genetic algorithms. In *ICGA*, volume 95, pages 184–192. Citeseer, 1995.

[58] E.T. Juengst, R.A. Settersten, J.R. Fishman, and M.L. McGowan. After the revolution? Ethical and social challenges in personalized genomic medicine. *Personalized Medicine*, 9(4):429–439, 2012.

[59] Leila Kallel, Bart Naudts, and Colin R Reeves. Properties of fitness functions and search landscapes. In *Theoretical aspects of evolutionary computing*, pages 175–206. Springer, 2001.

[60] Stuart Kauffman. *The origins of order: Self organization and selection in evolution*. Oxford University Press, 1993.

[61] Stuart Kauffman and Simon Levin. Towards a general theory of adaptive walks on rugged landscapes. *Journal of theoretical Biology*, 128(1):11–45, 1987.

[62] Stuart A Kauffman and Edward D Weinberger. The NK model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of theoretical biology*, 141(2):211–245, 1989.

[63] Eamonn Keogh and Shruti Kasetty. On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and knowledge discovery*, 7(4):349–371, 2003.

[64] Charles M Kilo. A framework for collaborative improvement: lessons from the institute for healthcare improvement's breakthrough series. *Quality Management in Healthcare*, 6(4):1–14, 1998.

[65] Bobby PC Koeleman, Benedicte Alexandre Lie, Dag Erik Undlien, Frank Dudbridge, Erik Thorsby, Rindert RP De Vries, Francesco Cucca, Bart O Roep, MJ Giphart, and John A Todd. Genotype effects and epistasis in type 1 diabetes and hla-dq trans dimer associations with disease. *Genes and immunity*, 5(5):381–388, 2004.

[66] J.S. Krupa, S. Chatterjee, E. Eldridge, D.M. Rizzo, and M.J. Eppstein. Evolutionary exploratory association discovery: A plug-in hybrid vehicle adoption application. *Submitted to the 21st International GECCO Confference*, 2012.

[67] J.A. LePine, R.F. Piccolo, C.L. Jackson, J.E. Mathieu, and J.R. Saul. A meta-analysis of teamwork processes: Tests of a multidimensional model and relationships with team effectiveness criteria. *Personnel Psychology*, 61(2):273–307, 2008.

[68] Rui Li, Michael TM Emmerich, Jeroen Eggermont, Ernst GP Bovenkamp, Thomas Bäck, Jouke Dijkstra, and Johan HC Reiber. Mixed-integer nk landscapes. In *Parallel Problem Solving from Nature-PPSN IX*, pages 42–51. Springer, 2006.

[69] Rung Tzuo Liaw and Chuan Kang Ting. Effect of model complexity for estimation of distribution algorithm in NK landscapes. In *2013 IEEE Symposium on Foundations of Computational Intelligence (FOCI)*, pages 76–83. IEEE, 2013.

[70] H. Liu, J. Li, and L. Wong. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Informatics Series*, pages 51–60, 2002.

[71] Rita Mangione-Smith, Matthias Schonlau, Kitty S Chan, Joan Keesey, Mayde Rosen, Thomas A Louis, and Emmett Keeler. Measuring the effectiveness of a collaborative for quality improvement in pediatric asthma care: does implementing the chronic care model improve processes and outcomes of care? *Ambulatory Pediatrics*, 5(2):75–82, 2005.

[72] T. Manser. Teamwork and patient safety in dynamic domains of healthcare: A review of the literature. *Acta Anaesthesiologica Scandinavica*, 53(2):143–151, 2008.

[73] Narine Manukyan, Margaret J Eppstein, and Jeffrey D Horbar. Team learning for healthcare quality improvement. *IEEE Access*, 1:545–557, 2013.

[74] Narine Manukyan, Margaret J Eppstein, and Donna M Rizzo. Data-driven cluster reinforcement and visualization in sparsely-matched self-organizing maps. *Neural Networks and Learning Systems, IEEE Transactions on*, 23(5):846–852, 2012.

[75] M.A. Marks, J.E. Mathieu, and S.J. Zaccaro. A temporally based framework and taxonomy of team processes. *Academy of Management Review*, pages 356–376, 2001.

[76] J.A. Martin, K.D. Kochanek, D.M. Strobino, B. Guyer, and M.F. MacDorman. Annual summary of vital statistics—2003. *Pediatrics*, 115(3):619, 2005.

[77] Klim McPherson, John E Wennberg, Ole B Hovind, Peter Clifford, et al. Small-area variations in the use of common surgical procedures: An international comparison of New England, England, and Norway. *The New England journal of medicine*, 307(21):1310, 1982.

[78] Brian S Mittman. Creating the evidence base for quality improvement collaboratives. *Annals of internal medicine*, 140(11):897–901, 2004.

[79] Naoki Miyagawa, Hiroshi Teramoto, Chun-Biu Li, and Tamiki Komatsuzaki. Decomposability of multivariate interactions. *Complex Systems*, 20(2):165, 2011.

[80] Douglas C Montgomery, Douglas C Montgomery, and Douglas C Montgomery. *Design and analysis of experiments*, volume 7. Wiley New York, 1984.

[81] Jason H Moore. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human heredity*, 56(1-3):73–82, 2003.

[82] Alberto Moraglio and Julian Togelius. Geometric differential evolution. In *Proceedings of the 11th Annual conference on Genetic and evolutionary computation*, pages 1705–1712. ACM, 2009.

[83] L.S. Morales, D. Staiger, J.D. Horbar, J. Carpenter, M. Kenny, J. Geppert, and J. Rogowski. Mortality among very low birthweight infants in hospitals serving minority populations. *American journal of public health*, 95(12):2206, 2005.

[84] Erum Nadeem, S Serene Olin, Laura Campbell Hill, Kimberly Eaton Hoagwood, and Sarah McCue Horwitz. Understanding the components of quality improvement collaboratives: A systematic literature review. *Milbank Quarterly*, 91(2):354–394, 2013.

[85] P.J. Newton, EJ Halcomb, PM Davidson, and A.R. Denniss. Barriers and facilitators to the implementation of the collaborative method: Reflections from a single site. *Quality and Safety in Health Care*, 16(6):409–414, 2007.

[86] I.S. Oh, J.S. Lee, and B.R. Moon. Hybrid genetic algorithms for feature selection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(11):1424–1437, 2004.

[87] World Health Organization et al. Who patient safety curriculum guide for medical schools. 2009.

[88] J. Øvretveit, P. Bate, P. Cleary, S. Cretin, D. Gustafson, K. McInnes, H. McLeod, T. Molfenter, P. Plsek, G. Robert, et al. Quality collaboratives: lessons from research. *Quality and safety in health care*, 11(4):345–351, 2002.

[89] J Øvretveit, Paul Bate, Paul Cleary, Shan Cretin, D Gustafson, K McInnes, H McLeod, T Molfenter, P Plsek, Glenn Robert, et al. Quality collaboratives: lessons from research. *Quality and safety in health care*, 11(4):345–351, 2002.

[90] Ray Pawson and Nick Tilley. *Realistic evaluation*. Sage, 1997.

[91] N.R. Payne, M.J. Finkelstein, M. Liu, J.W. Kaempf, P.J. Sharek, and S. Olsen. Nicu practices and outcomes associated with 9 years of quality improvement collaboratives. *Pediatrics*, 125(3):437–446, 2010.

[92] Martin Pelikan. Analysis of estimation of distribution algorithms and genetic algorithms on NK landscapes. In *Proceedings of the 10th annual conference on Genetic and evolutionary computation*, pages 1033–1040. ACM, 2008.

[93] F. Pernkopf and P. O'Leary. Feature selection for classification using genetic algorithms with a novel encoding. In *Computer Analysis of Images and Patterns*, pages 161–168. Springer, 2001.

[94] Charles Perrow. *Normal Accidents: Living with High Risk Technologies (Updated)*. Princeton University Press, 2011.

[95] Paul E Plsek. Collaborating across organizational boundaries to improve the quality of care. *American journal of infection control*, 25(2):85–95, 1997.

[96] Paul E Plsek and Trisha Greenhalgh. The challenge of complexity in health care. *Bmj*, 323(7313):625–628, 2001.

[97] Paul E Plsek and Tim Wilson. Complexity, leadership, and management in healthcare organisations. *Bmj*, 323(7315):746–749, 2001.

[98] P.E. Plsek. Collaborating across organizational boundaries to improve the quality of care. *American journal of infection control*, 25(2):85–95, 1997.

[99] B.C. Poulton and M.A. West. Effective multidisciplinary teamwork in primary health care. *Journal of Advanced Nursing*, 18(6):918–925, 2008.

[100] M.L. Raymer, W.F. Punch, E.D. Goodman, L.A. Kuhn, and A.K. Jain. Dimensionality reduction using genetic algorithms. *Evolutionary Computation, IEEE Transactions on*, 4(2):164–171, 2000.

[101] Colin Reeves and Christine Wright. An experimental design perspective on genetic algorithms. In *Foundations of Genetic Algorithms 3*, 1995.

[102] Colin R Reeves. Experiments with tuneable fitness landscapes. In *Parallel Problem Solving from Nature PPSN VI*, pages 139–148. Springer, 2000.

[103] Colin R Reeves and Christine C Wright. Epistasis in genetic algorithms: An experimental design perspective. In *Proceedings of the 6th International Conference on Genetic Algorithms*, pages 217–224. Morgan Kaufmann Publishers Inc., 1995.

[104] Ian Reid. Complexity science: Let them eat complexity: the emperor's new toolkit. *BMJ: British Medical Journal*, 324(7330):171, 2002.

[105] D.N. Reshef, Y.A. Reshef, H.K. Finucane, S.R. Grossman, G. McVean, P.J. Turnbaugh, E.S. Lander, M. Mitzenmacher, and P.C. Sabeti. Detecting novel associations in large data sets. *science*, 334(6062):1518–1524, 2011.

[106] A.D. Rodrigues. *Drug-drug interactions*. Informa Healthcare, New York, NY, 2008.

[107] J.A. Rogowski, J.D. Horbar, P.E. Plsek, L.S. Baker, J. Deterding, W.H. Edwards, J. Hocker, A.D. Kantak, P. Lewallen, W. Lewis, et al. Economic implications of neonatal intensive care unit collaborative quality improvement. *Pediatrics*, 107(1):23, 2001.

[108] J.A. Rogowski, J.D. Horbar, D.O. Staiger, M. Kenny, J. Carpenter, and J. Geppert. Indirect vs direct hospital quality indicators for very low-birth-weight infants. *JAMA: the journal of the American Medical Association*, 291(2):202, 2004.

[109] J.A. Rogowski, D.O. Staiger, and J.D. Horbar. Variations in the quality of care for very-low-birthweight infants: implications for policy. *Health Affairs*, 23(5):88–97, 2004.

[110] Jani Rönkkönen, Xiaodong Li, Ville Kyrki, and Jouni Lampinen. A framework for generating tunable test functions for multimodal optimization. *Soft Computing*, 15(9):1689–1706, 2011.

[111] Peter M Rothwell. External validity of randomised controlled trials:to whom do the results of this trial apply?. *The Lancet*, 365(9453):82–93, 2005.

[112] William Rowe, Mark Platt, David C Wedge, Philip J Day, Douglas B Kell, and Joshua Knowles. Analysis of a complete dna–protein affinity landscape. *Journal of The Royal Society Interface*, 7(44):397–408, 2010.

[113] Bill Runciman and Merrilyn Walton. *Safety and ethics in healthcare: a guide to getting it right*. Ashgate Publishing, Ltd., 2007.

[114] E. Salas, N.J. Cooke, and M.A. Rosen. On teams, teamwork, and team performance: Discoveries and developments. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 50(3):540–547, 2008.

[115] T. Sandmann and M. Boutros. Screens, maps & networks: From genome sequences to personalized medicine. *Current Opinion in Genetics & Development*, 22:36–44, 2012.

[116] Elad Schneidman, Susanne Still, Michael J Berry, William Bialek, et al. Network information and connected correlations. *Physical review letters*, 91(23):238701, 2003.

[117] Matthias Schonlau, Rita Mangione-Smith, Kitty S Chan, Joan Keesey, Mayde Rosen, Thomas A Louis, Shin-Yi Wu, and Emmett Keeler. Evaluation of a quality improvement collaborative in asthma care: does it improve processes and outcomes of care? *The Annals of Family Medicine*, 3(3):200–208, 2005.

[118] L.M.T. Schouten, R.P.T.M. Grol, and M.E.J.L. Hulscher. Factors influencing success in quality-improvement collaboratives: Development and psychometric testing of an instrument. *Implementation Science*, 5(1):1–9, 2010.

[119] L.M.T. Schouten, M.E.J.L. Hulscher, J.J.E. Everdingen, R. Huijsman, and R.P.T.M. Grol. Evidence for the impact of quality improvement collaboratives: Systematic review. *Bmj*, 336(7659):1491–1494, 2008.

[120] Loes MT Schouten, MEJL Hulscher, Jannes JE Van Everdingen, Robbert Huijsman, Louis W Niessen, and RPTM Grol. Short-and long-term effects of a quality improvement collaborative on diabetes management. *Implement Sci*, 5:94, 2010.

[121] Paul G Shekelle, Peter J Pronovost, Robert M Wachter, Stephanie L Taylor, Sydney M Dy, Robbie Foy, Susanne Hempel, Kathryn M McDonald, John Ovretveit, Lisa V Rubenstein, et al. Advancing the science of patient safety. *Annals of Internal Medicine*, 154(10):693–696, 2011.

[122] Stephen M Shortell, Jill A Marsteller, Michael Lin, Marjorie L Pearson, Shin-Yi Wu, Peter Mendel, Shan Cretin, and Mayde Rosen. The role of perceived team effectiveness in improving chronic illness care. *Medical care*, 42(11):1040–1048, 2004.

[123] Leif I Solberg. If youve seen one quality improvement collaborative. *The Annals of Family Medicine*, 3(3):198–199, 2005.

[124] Kenneth Tan, Gordon Baxter, Simon Newell, Steve Smye, Peter Dear, Keith Brownlee, and Jonathan Darling. Knowledge elicitation for validation of a neonatal ventilation expert system utilising modified delphi and focus group techniques. *International journal of human-computer studies*, 68(6):344–354, 2010.

[125] Reiko Tanese. *Distributed genetic algorithms for Function Optimization*. PhD thesis, The University of Michigan, Ann Arbor, MI, 1989.

[126] S.L. Taylor, S. Dy, R. Foy, et al. What context features might be important determinants of the effectiveness of patient safety practice interventions? *BMJ Quality & Safety*, 20(7):611–617, 2011.

[127] Dirk Thierens. The linkage tree genetic algorithm. In *Parallel Problem Solving from Nature, PPSN XI*, pages 264–273. Springer, 2010.

[128] Dirk Thierens and Peter AN Bosman. Hierarchical problem solving with the linkage tree genetic algorithm. In *Proceeding of the fifteenth annual conference on Genetic and evolutionary computation conference*, pages 877–884. ACM, 2013.

[129] Nicholas Tomko, Inman Harvey, and Andrew Philippides. Unconstrain the population: The benefits of horizontal gene transfer in genetic algorithms. In *SmartData*, pages 117–127. Springer, 2013.

[130] Shaun Treweek and Merrick Zwarenstein. Making trials matter: pragmatic and explanatory trials and the problem of applicability. *Trials*, 10(37):9, 2009.

[131] M.E. Turner. *Groups at work: Theory and research*. Lawrence Erlbaum, Hillsdale, NJ, 2000.

[132] Ryan J Urbanowicz and Jason H Moore. The application of michigan-style learning classifiersystems to address genetic heterogeneity and epistasisin association studies. In *Proceedings of the 12th annual conference on Genetic and evolutionary computation*, pages 195–202. ACM, 2010.

[133] G. Vaidyanathan. Redefining clinical trials: The age of personalized medicine. *Cell*, 148(6):1079–1080, 2012.

[134] Vesselin K Vassilev, Terence C Fogarty, and Julian F Miller. Information characteristics and the structure of landscapes. *Evolutionary Computation*, 8(1):31–60, 2000.

[135] Edward Weinberger. Correlated and uncorrelated fitness landscapes and how to tell the difference. *Biological cybernetics*, 63(5):325–336, 1990.

[136] John Wennberg and Alan Gittelsohn. Small area variations in health care delivery: A population-based health information system can guide planning and regulatory decision-making. *Science*, 182(4117):1102–1108, 1973.

[137] John E Wennberg. *Tracking Medicine: A Researcher's Quest to Understand Health Care*. Oxford University Press, USA, 2010.

[138] M.A. West. *Effective teamwork: Practical lessons from organizational research*. Blackwell Publishing, Oxford, 2012.

[139] Tim Wilson, Donald M Berwick, and Paul D Cleary. What do collaborative improvement projects do? experience from seven countries. *Joint Commission Journal on Quality and Patient Safety*, 29(2):85–93, 2003.

[140] Tim Wilson, Tim Holt, and Trisha Greenhalgh. Complexity and clinical care. *Bmj*, 323(7314):685–688, 2001.

[141] David D Woods, Leila J Johannesen, Richard I Cook, and Nadine B Sarter. Behind human error: Cognitive systems, computers and hindsight. Technical report, DTIC Document, 1994.

[142] David D Woods, Emily S Patterson, and Richard I Cook. Behind human error: taming complexity to improve patient safety. *Handbook of Human Factors and Ergonomics in Health Care and Patient Safety. London: Lawrence Erlbaum*, pages 459–76, 2007.

[143] Michael Wooldridge. *An introduction to multiagent systems*. John Wiley & Sons, 2009.

[144] Alden H Wright, Richard K Thompson, and Jian Zhang. The computational complexity of NK fitness functions. *IEEE Transactions on Evolutionary Computation*, 4(4):373–379, 2000.

[145] S. Wright. The roles of mutation, inbreeding, crossbreeding and selection in evolution. In *Proceedings of the sixth international congress on genetics*, volume 1, pages 356–366, 1932.

[146] Paul C Young, Gordon B Glade, Gregory J Stoddard, and Chuck Norlin. Evaluation of a learning collaborative to improve the delivery of preventive services by pediatric practices. *Pediatrics*, 117(5):1469–1476, 2006.

[147] Z.J. Yu, F. Haghighat, B. Fung, and L. Zhou. A novel methodology for knowledge discovery through mining associations between building operational data. *Energy and Buildings*, 2011.

[148] J.A.F. Zupancic, D.K. Richardson, J.D. Horbar, J.H. Carpenter, S.K. Lee, G.J. Escobar, et al. Revalidation of the score for neonatal acute physiology in the vermont oxford network. *Pediatrics*, 119(1):e156–e163, 2007.

Chapter 3

Team Learning for Healthcare Quality Improvement

Manukyan, N., Eppstein, M.J., and Horbar. J. “Team Learning for Healthcare Quality Improvement”, IEEE Access, 1:545-557, 2013.

3.1 Abstract

In organized healthcare quality improvement collaboratives (QICs), teams of practitioners from different hospitals exchange information on clinical practices, with the aim of improving health outcomes at their own institutions. However, what works in one hospital may not work in others with different local contexts, due to non-linear interactions among various demographics, treatments, and practices. In previous studies of collaborations where the goal is collective problem solving, teams of diverse individuals have been shown to outperform teams of similar individuals. However, when the purpose of collaboration is knowledge diffusion in complex environments, it is not clear whether team diversity will

help or hinder effective learning. In this study, we first use an agent-based model of QICs to show that teams comprising similar individuals outperform those with more diverse individuals under nearly all conditions, and that this advantage increases with the complexity of the landscape and the level of noise in assessing performance. Examination of data from a network of real hospitals provides encouraging evidence of a high degree of similarity in clinical practices, especially within teams of hospitals engaging in QIC teams. However, our model also suggests that groups of similar hospitals could benefit from larger teams and more open sharing of details on clinical outcomes than is currently the norm. To facilitate this, we propose a secure virtual collaboration system that would allow hospitals to efficiently identify potentially better practices in use at other institutions similar to theirs, without any institutions having to sacrifice the privacy of their own data. Our results may also have implications for other types of data-driven diffusive learning, such as in personalized medicine and evolutionary search in noisy, complex combinatorial optimization problems.

3.2 Introduction

Much research has focused on studying team collaboration for joint problem solving [17, 67, 72, 75, 99, 114, 131, 138]. In this context, diverse teams have been shown to offer some advantage. For example, in [41] the authors show that groups of diverse problem solvers can outperform more homogeneous groups of higher-ability problem solvers, because diverse individuals bring different perspectives and heuristics that aid in the creativity of the collective intelligence. Similarly, in [39] the authors show that teams with higher numbers of newcomers perform better because newcomers add to the diversity of the team. On the other hand, in personalized recommendation systems with collaborative filtering, historical data of users with similar preferences are used for making personalized recommendations [33, 34]. However, when the purpose of team collaboration is knowledge diffusion in complex environments, rather than collective problem solving or inference of preferences, it is not clear whether diverse teams help or hinder performance improvement of the individual team members.

Many clinicians are now participating in organized quality improvement collaboratives (QICs), in which teams of practitioners from different healthcare organizations exchange information on selected clinical practices and outcomes. Nonprofit institutions such as the Vermont Oxford Network (VON) [42, 52] act as facilitators for these QICs. Team members identify potentially better practices in use at teammates' institutions and then try them out in the local context of their home institutions [12, 49]. In this type of collaborative environment, the goal is for all hospitals to improve their own performance by learning from the experiences of others in their teams. However, what works in one hospital might not work in others with different local contexts, due to non-linear interactions among various treatments and practices. Indeed, it is becoming increasingly recognized that such complex interactions are not uncommon in healthcare [18, 37, 54, 56, 106]. While there is positive but limited evidence that QICs can result in improved quality of care [84, 119], it is not clear which factors contribute to the effectiveness of teamwork in QICs [85, 118, 126]. The primary goal of this contribution is to advance our understanding of how different strategies of team formation are likely to affect quality improvement in healthcare through information sharing and learning.

In [27], we developed an agent-based model (ABM) where agents represent healthcare institutions searching for combinations of clinical practices that improve the health outcomes of their patients. In that work, we showed that simulated multi-institutional QICs often perform better than simulated randomized controlled trials, due to a combination of greater statistical power and more context-dependent evaluation of practices, especially in noisy, complex environments with multiple interactions between clinical practices. We also showed that search was improved when the hospitals were "clustered" (rather than uniformly randomly "scattered") in the landscape of clinical practices, and argued that real hospitals were more likely to be clustered based on their long history of information sharing. Interestingly, we found that initially clustered agents actually became more diverse after searching together through repeated QICs. However, in [27], team members were randomly selected for each set of trials, team sizes were held constant, no data on real hospitals was provided to support the assumption regarding clustering, and no explanation was provided for why populations of clustered agents became increasingly diverse as their fitness improved.

Here, we use a similar ABM to study the interacting impacts of various aspects of team formation on individual performance improvement and diversity in QIC teams. We performed a preliminary analysis of real data from hospitals participating in QICs showing that these hospitals are, indeed, clustered. Based on this, we developed a new method for clustered initialization of synthetic agents, such that the distribution of distances between agent attributes resembles the observed hospital distribution. We also developed an $O(n \log n)$ approximation algorithm to the NP-hard problem of creating equal-sized teams of n individuals, each team with maximum within-team similarity. We then assessed the sensitivity of performance improvement to a variety of factors, including (i) within-team diversity, (ii) frequency of team reformation, (iii) clustered vs. scattered initial populations, and (iv) how often hospitals should wait before being allowed to reevaluate the same practice. The impacts of these factors are studied under a variety of scenarios, with varying degrees of noise in fitness evaluation and number of two-feature interactions between practices. Finally, we analyzed a larger set of hospital data to try to assess how the characteristics of real QICs relate to those found to be best in our ABM. Based on our study, we propose potential ways to facilitate learning in healthcare environments and other domains.

This paper is organized as follows: In Section 3.3, we describe the methods used in the ABM portion of the study. In Section 3.4, we show the results of the ABM portion of the study. Section 3.5 discusses data curation and analysis of real hospital data. In Section 3.6, we show the results of the real hospital data analysis. Finally, in Sections 3.7 and Section 3.8 we provide discussion and conclusions.

3.3 Methods

3.3.1 Modeling the Problem

We use the same clinical fitness landscape model as used in [27], where hospitals are modeled as agents trying to find sets of clinical practices that improve health outcomes for their patient population. The probability of patient survival $Pr(s_x)$ (or some other desired outcome) at a given healthcare institution is simulated with a high dimensional logistic function as follows:

$$Pr(s_x) = \Bigg(1 + exp\Bigg(-\Bigg(\beta_0 + \sum_{i=1}^{N} \beta_i x_i + \sum_{i=1}^{(N-1)} \sum_{j=i+1}^{N} \gamma_{ij} x_i x_j + H\Bigg)\Bigg)\Bigg)^{-1} \tag{3.1}$$

where x is a vector of N binary features ($x_i \in \{-1, 1\}$), each representing the presence or absence of the use of a specific practice, intervention, or other modifiable characteristic of the institution. Coefficients β_i and γ_{ij} are randomly drawn from a normal distribution with a mean of 0 and standard deviation of $L^{-0.5}$, where L is the total nuber of non-zero terms in the model. As in [27] we restrict our landscapes to those with an average fitness of 0.5 ($\beta_0 = 0$), include non-zero coefficients (β_i) for all main effects, and only model up to two-feature interactions (γ_{ij}); i.e., potential higher order interactions (H) are always set to zero. In a noise-free environment we calculate the probability of patient survival using Eq. (1). To model heterogeneity in patient-level responses we use Bernoulli trials with survival probability given by Eq. (1). Thus, trials with fewer patients have higher levels of noise in the fitness function, due to stochastic effects. In the remainder of this manuscript, we use the terms "agent" and "individual" to mean an abstraction of a healthcare institution.

3.3.2 Population Initialization

In [27] we compared search strategies starting from initially scattered or clustered initial populations of agents on landscapes of simulated clinical practices, and argued that the latter was more realistic. In the first case, uniformly scattered populations of M agents were created with N randomly generated binary features. The expected median pairwise normalized Hamming distances (nHDs) in scattered populations is 0.5. However, with no real data to guide us at the time, the clustered populations in [27] were generated by simply starting with a population of identical copies of a random individual and perturbing random features until the desired median nHD of 0.1 was achieved, resulting in an N-dimensional roughly spheroidal cluster of binary vectors. For the current study, we first analyzed self-reported data from a VON survey on 93 binarized practice values from 51 VON hospitals, each participating in at least one of 7 VON-sponsored QIC teams that met in September 2003. These data showed a median pairwise nHD of only 0.34, ranging from 0 to 0.73 (Fig. 3.1a). Although these observations are limited, they do support the notion that hospitals are clustered rather than scattered in the feature space, although not to the degree modeled in [27]. We used this data to guide the development of a new algorithm for clustered population initialization that we refer to as $MakeSnakingCluster$, which can generate clustered distributions more similar to that of the observed hospitals. As in [27], we compare search results between populations with initially clustered and scattered distributions. (Note: in section 3.5 we show additional analysis of a larger data set of 20 real-valued clinical practices that further supports the notion of clustered hospitals.)

The $MakeSnakingCluster$ algorithm for binary-featured landscapes works as follows. There are two tunable parameters that control the resulting distribution; d is a specified Hamming distance (HD), and K is an integer between 1 and $M - 1$, where M is the number of agents. First we create a random individual with N binary features as the core individual. Then we create K individuals that are d HD away from the core individual by flipping d randomly selected bits of each of K copies of the core individual. We next pick one of these K generated individuals as the core individual for the next step and repeat the process. The algorithm terminates when M individuals have been generated (if M is not evenly divisible by K the last iteration is terminated early).

Although we define and use the algorithm above for binary-featured landscapes, it can be generalized to work for landscapes with real-valued features by replacing the HDs with Euclidean distances (EDs). In Fig. 3.2 we illustrate an example population generated by the $MakeSnakingCluster$ algorithm in a 2-dimensional real-valued feature space, since this is easier to visualize than an N-dimensional binary space. Notice that the $MakeSnakingCluster$ algorithm generates a non-spheroidal cluster of individuals that tends to snake through the landscape, hence the name.

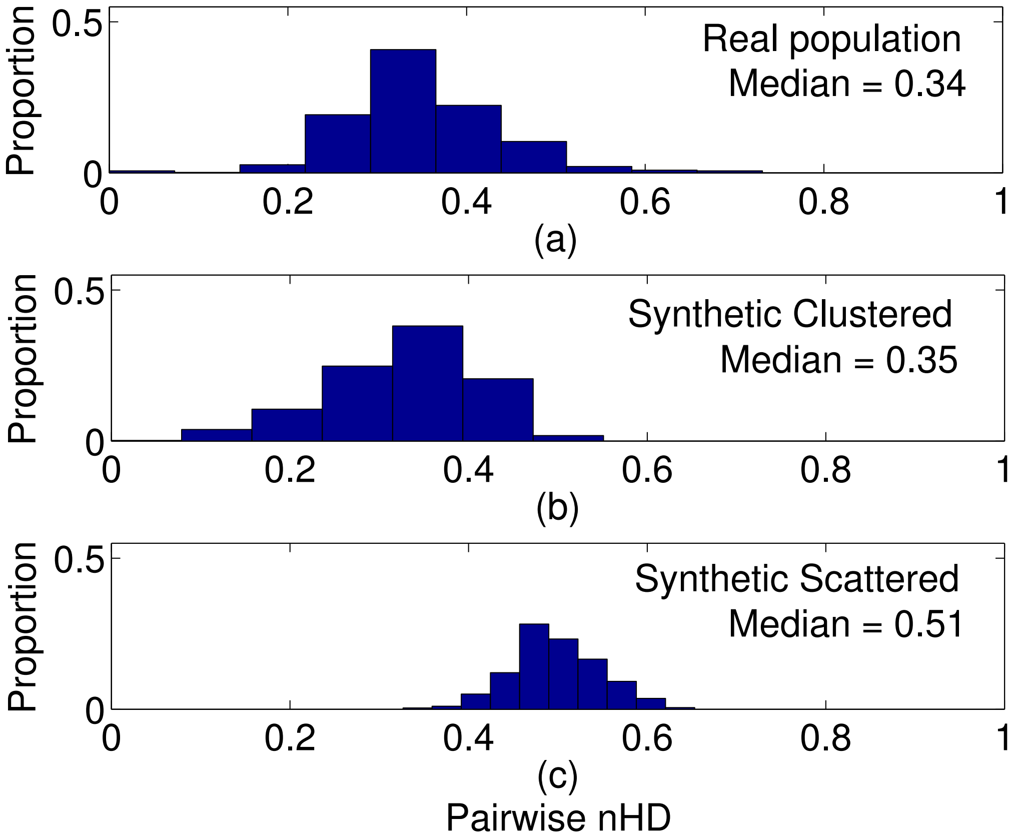
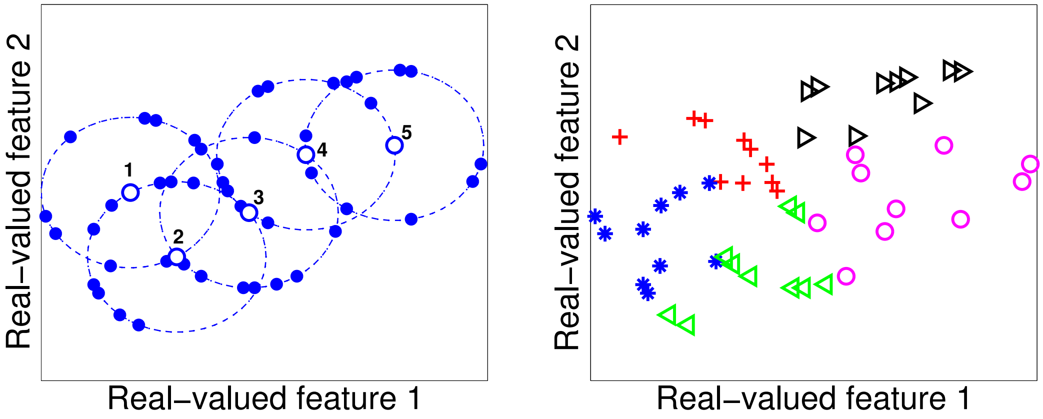


Figure 3.1: Representative histograms of proportions of pairwise nHDs of $M = 51$ agents, each with $N = 93$ features, for a) a dataset of real hospitals with binarized practices as features, b) clustered synthetic random agents with binary features, generated by $MakeSnakingCluster$ with $K = 10$ and $d = 13$, c) scattered synthetic random agents with binary features.



(a)

(b)

Figure 3.2: a) Illustration of one instance of a population created with the $MakeSnakingCluster$ algorithm in 2-D real-valued feature space ($M = 50$, $K = 10$ and $d = 13$), where numbered open circles represent core individuals in each step. b) Illustration of the population shown in a, divided into $T = 5$ teams picked by the $PickSimilarTeams$ algorithm, where each team is shown by a unique color and shape combination.

We compare the distribution of all pairwise HDs of the single instance of hospital data described above with that of a single instance of a clustered population generated by the $MakeSnakingCluster$ algorithm to create $M = 51$ individuals with $N = 93$ features, where we tuned $K = 10$ and $d = 13$ to achieve a median pairwise nHD (Fig. 3.1b) that is close to that of the real hospital data (Fig. 3.1a). For the remainder of our ABM simulations, we used $K = 10$ and $d = 13$ to generate random clustered populations. Note that the distribution of one instance of a scattered population with the same N and M (Fig. 3.1c) is very different from that of the real hospitals.

3.3.3 Team Structure

One potentially important influence on team learning is the team construction mechanism; i.e., deciding which agents should be in the same teams. In our ABM we compare randomly formed teams (as used in [27]) to teams formed by the principal of homophily, in which similar agents are grouped together. Since picking equal-sized teams with maximum within-team similarity is an NP-hard problem, we devised the following $O(n \log n)$ approximation algorithm we call $PickSimilarTeams$.

To place M agents into T teams of M_T homophilous agents (where $M_T = \lceil \frac{M}{T} \rceil$), we first calculate all the pairwise HDs in the population. Then for each agent we calculate the mean of the HDs between the agent and its most similar $M_T - 1$ neighbors in the population. The first team is selected to be the agent with the smallest calculated mean HD to its $M_T - 1$ closest neighbors. We then remove the individuals that were assigned to this team from the available population and repeat the process for the remaining population until we have T teams. (Note that the first team picked by this approximation algorithm will have maximum within-team similarity, but that teams picked later may have lower within-team simiilarity, so the resulting teams are not necessarily optimally homophilous.)

A visual illustration of the $PickSimilarTeams$ algorithm is shown in Fig. 3.2b, where the algorithm has divided the population of $M = 50$ individuals shown in Fig. 3.2a into T = 5 teams, with $M_T = 10$ individuals in each team.

3.3.4 Team Learning

We use the team learning algorithm described in [27], with minor modifications. In each generation, each agent selects one feature that has the highest difference between the average feature value of its teammates that have higher and lower fitnesses than the agent, and such that the selected feature of the agent is different from the majority feature value of the fitter teammates. The agent then flips the bit for this feature and tries this new feature combination (calculates the fitness) in its local context and, if it is better than the previous feature combination, it adopts the new feature value. Unlike in [27], where the most fit member of each team does no exploration, in this study the fittest individual in each team selects the feature that has the smallest difference between the agent's feature value and the average of all other teammates' feature values, tries the complement of its feature value, and adopts it if better. Agents are not allowed to retry the same features within $tabu$ trial steps. In [27] we used $tabu = 1$, but in this study we experiment with a range of $tabu$ values.

The feature selection strategy we describe above can mitigate the effects of noise in fitness evaluation (as does the approach in [21]), while also providing agent-specific customized recommendations for change based on where each agent's fitness lies relative to the others in its team.

3.3.5 Simulations

In all simulations reported here, we assessed the impact and interactions of different factors on performance improvement through team learning for $M = 100$ agents (representing hospitals), each with $N = 100$ binary-valued features (representing clinical practices). Specifically, we varied the initial population type, the team formation mechanism, team size (M_T), the number of trial steps between team reformation, the amount of noise in fitness evaluation, the number of two-feature interactions included in the fitness function (Eq. (1)), and the length of the $tabu$ period, as shown in Table 3.1. Unless otherwise specified, we used the default values of $M_T = 10$, with team reformation after each trial step, and $tabu = 5$ (default values are shown in bold in Table 3.1).

Table 3.1: Factors varied in the ABM simulations.

Factor	Values Studied
Initial Population Type	{Clustered, Scattered}
Team Formation Mechanism	{Homophilous, Random}
Team Size (M_T)	{1, 2, 4, 5, **10**, 20, 25, 50, 100}
Steps between Team Reformation	{**1**, 4, 6, 10, 20, 50, ∞ }

Noise in Fitness Evaluation
(corresponding # patients per trial)
	{None, Low, High}
	{(∞), (320), (40) }

two-feature interactions
(% possible two-feature interactions)
	{0, 495, 2475}
	{(0%), (10%), (50%)}

tabu	{2, **5**, 10} steps

Note that the degree of clustering in the initial population and whether teams are selected randomly or homophilously both affect the initial degree of within-team similarity, as shown in Table 3.2. We define "within-population nHD" as the mean of all pairwise normalized Hamming distances (nHDs) in the entire population (normalized by dividing by N). We define "within-team nHD" as the mean of the mean of the pairwise nHDs within each team. The abbreviations shown in Table 3.2 are used to label subsequent plots.

Other factors also interact to affect the changing degree of within-team similarity during the search process. For example, team reformation can either increase or decrease within-team similarity based on whether teams are formed homophilously or randomly. Noise and landscape complexity tend to increase inter-agent diversity as learning progresses, due to stochastic effects and the presence of multiple peaks in the landscape due to feature interactions, respectively.

Table 3.2: Average nHDs in 100 instances of initial populations and teams, for each of the four combinations of initial population type and team formation mechanism. Lower nHDs mean greater similarity.

Abbr.	Init Population	Team Formation	Within-Team nHD	Within-Pop. nHD
CH	Clustered	Homophilous	0.21	0.35
CR	Clustered	Random	0.33	0.35
SH	Scattered	Homophilous	0.47	0.50
SR	Scattered	Random	0.50	0.50

We generated 100 random landscapes for each specified number of two-feature interactions using Eq. (1) and generated one clustered and one scattered initial population for each landscape. All experiments with a given combination of parameter settings were averaged over the performances on these same 100 landscapes, starting from the same scattered or clustered populations, for 100 trial steps.

3.3.6 Statistical Comparisons

Pairs of experiments that differed in only one parameter were statistically compared as follows. We integrated each fitness curve over all 100 trial steps, for each of the 100 random landscapes with the specified number of two-feature interactions. We compared these integrated values using 2-tailed paired t-tests to asses for statistically signficant differences.

3.4 Results

Team search consistently significantly outperformed random search ($p < 0.01$, Fig. 3.3), consistent with [27], although in that study only randomly formed teams were studied. When averaged over 100 random landscapes, the performance of individual random

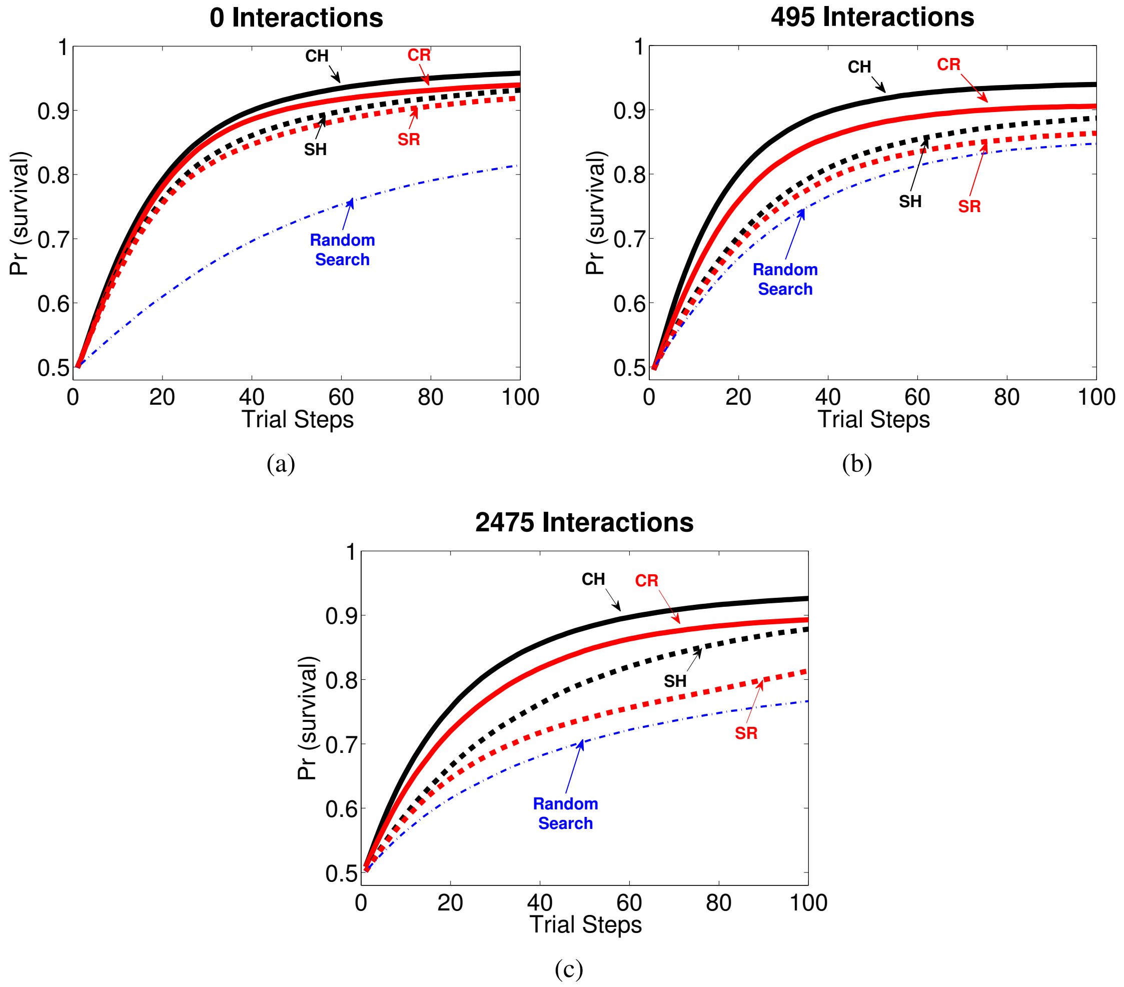
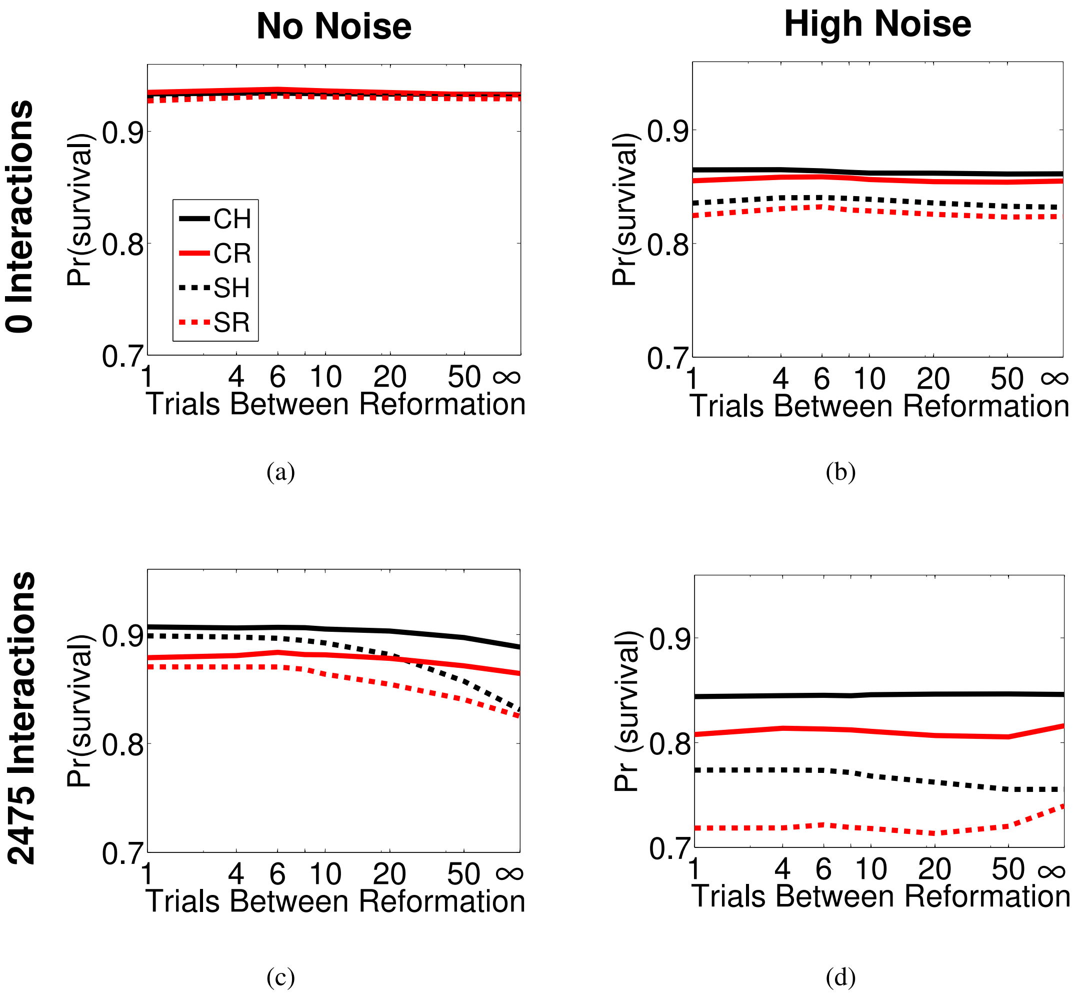


Figure 3.3: Mean probability of patient survival on 100 random landscapes at each of 100 trial steps, using 40 patients per trial on landscapes with a) 0, b) 495, and c) 2475 two-feature interactions.

searchers was statistically indistinguishable whether started from initially scattered or initially clustered populations, so we only show one of these curves in Fig. 3. This finding indicates that there is no inherent fitness advantage conferred by either of these two types of population initializations.

Our results also show that, at least when fitness evaluation is noisy (as is to be expected when hospitals try out a new practice on a relatively few patients in their own institution) the more internally similar the teams were, the better they performed (Fig. 3.3, with 40 patients per trial). Note that the order of performance from highest to lowest fitness shown in Fig. 3.3 matches with the order of initial within-team similarity shown in Table 3.2, with performance order being CH $>$ CR $>$ SH $>$ SR (each relation significant at the $p < 0.01$ level). This implies that agents are more effectively learning from teammates that have similar local contexts, and it can be seen that the relative advantage of less diverse teams increases as the complexity of the landscape increases (compare Fig. 3.3a,b,c).

In [27], teams were reformed between every trial step. To understand the impact of frequency of team reformation in simulated collaborations, we varied the frequency with which teams were reformed. In general, our results show that team search is relatively insensitive to the frequency of team reformation, especially when starting from clustered landscapes (Fig. 3.4). Homophilous teams of clustered agents (Fig. 3.4, black lines) were the least sensitive to frequency of team reformation, because there is relatively little switching of agents between teams even after they are reformed. This combination also consistently outperformed the strategies with more diverse teams, both for different levels of noise (compare Fig. 3.4a,c with no noise to 3.4b,d with high noise) and landscape complexity (compare Fig. 3.4a,b with no two-feature interactions to 3.4c,d with 2475 two-feature interactions). One apparent anomaly in these results occurs on complex



(a)

(b)

(c)

(d)

Figure 3.4: Mean probability of patient survival over 100 trial steps, averaged over 100 random landscapes, shown as a function of the frequency of team reformation. a) No two-feature interactions in the fitness landscapes and no noise in trials, b) No two-feature interactions in the fitness landscapes and noise in trials (40 patients per trial), c) 2475 two-feature interactions in the fitness landscapes and no noise in trials, and d) 2475 two-feature interactions in the fitness landscapes and noise in trials (40 patients per trial).

landscapes with no noise (Fig. 3.4c). Here, we observe a qualitative switch in the relative performance between the SH and CR combinations as the number of trial steps between team formation increases. Closer examination revealed that this occurs because frequent homophilous team reformation actually enables initially scattered populations to become highly clustered as agents converge towards each other in noise-free learning, ultimately achieving lower within-team nHDs than those that start initially clustered but are subject to frequent random team reformation. When fitness evaluation is noisy and landscapes are complex, there is actually a small but detectable increase in the performance of randomly formed teams as the number of trial steps between team reformations increases above 20 (Fig. 3.4d), since team members that stick together longer finally begin to converge towards each other, thereby promoting learning from more similar teammates.

Thus, the act of team reformation can have different influences on learning rate, depending on the direction and degree of its influence on within-team similarity. To illustrate this, consider a complex landscape (2475 two-feature interactions) and an initially scattered population, with teams formed prior to the initial trial and then not reformed again until trial step 50. After 50 steps of noise-free learning within the same teams of 10 agents, learning stagnates (Fig. 3.5a, black line before reformation at trial step 50); at this point, reformation into more homophilous teams causes an abrupt drop in within-team nHD (Fig. 3.5b, black line at 50 trial steps) with a consequent jump in the rate of fitness improvement (Fig. 3.5a, black line after reformation at trial step 50). On the other hand, when fitness evaluation is noisy it takes longer for team members to converge, so learning is slower (Fig. 3.5a, red line before reformation at trial step 50) and within-team nHD is still high even after 50 trials steps in the same team; at this point, a random reshuffling of team members causes an abrupt rise in within-team nHD (Fig. 3.5b, red line at trial step 50) and the rate of learning decreases even more (Fig. 3.5a, red line after reformation at trial step 50).

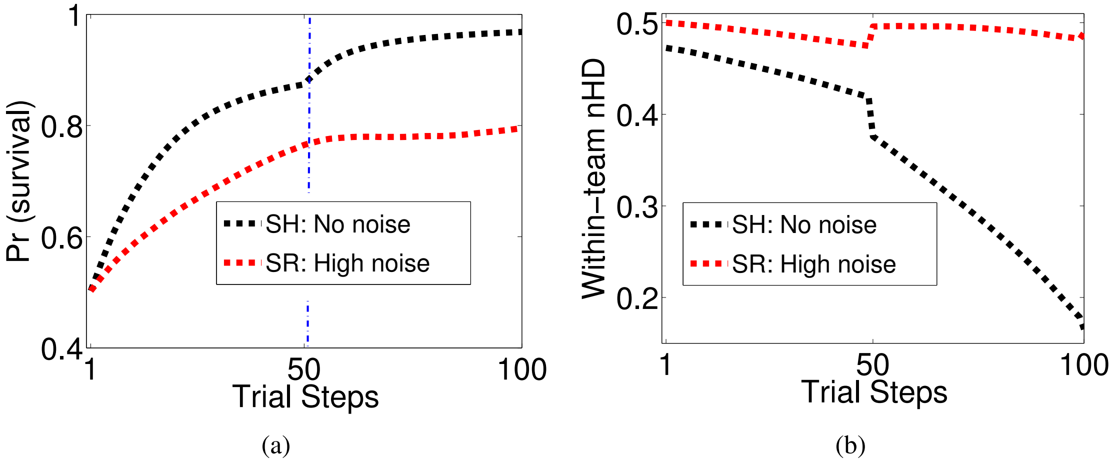
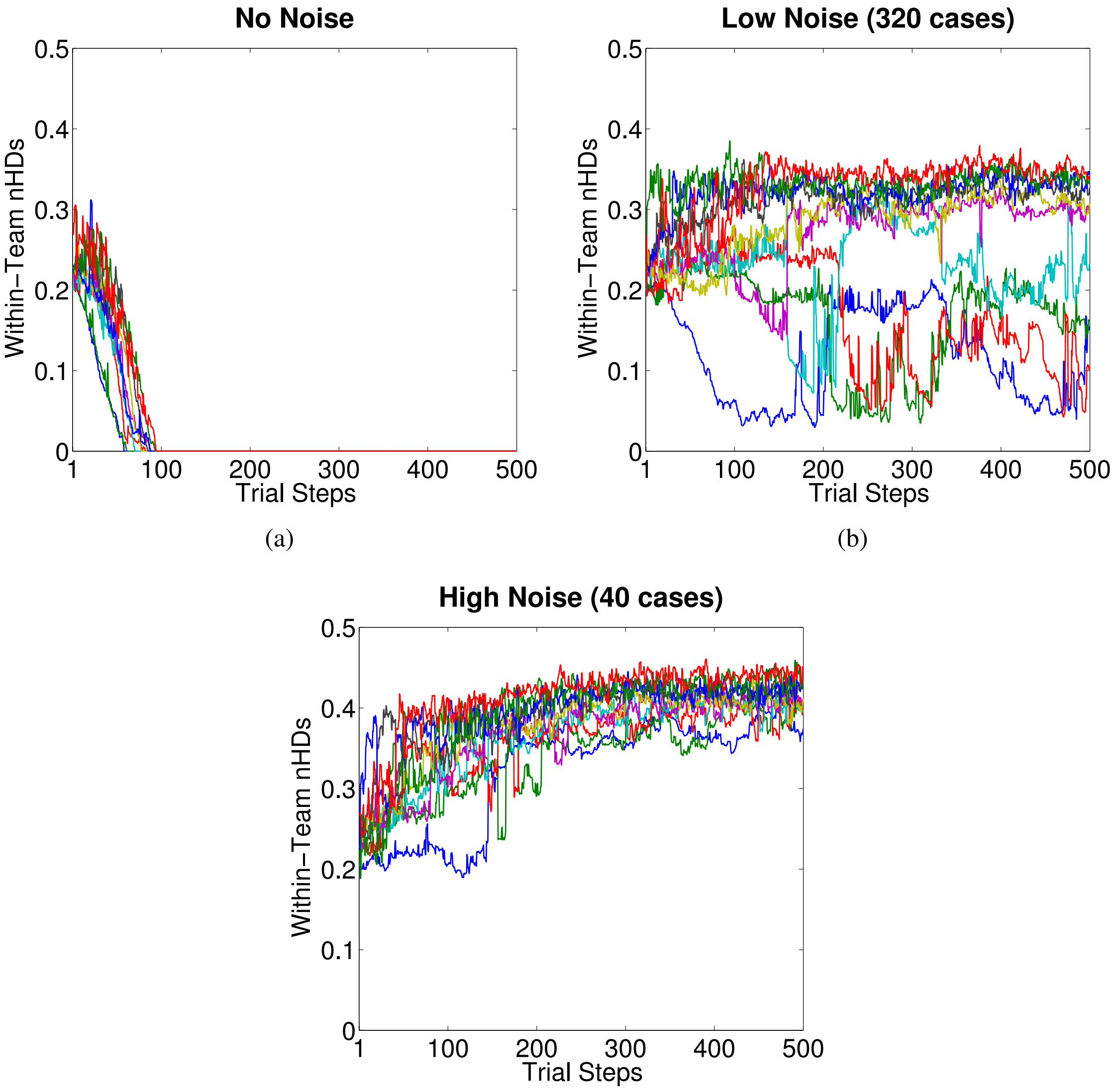


Figure 3.5: The effect of a single team reformation at trial step 50, starting from an initially scattered population on a complex landscape with 2475 two-feature interactions when fitness evaluation is noise-free and the team reformation is homophilous (black lines) or when fitness evaluation is noisy (only 40 patients per trial) and team reformation is random (red lines). a) Mean probability of patient survival on 100 random landscapes; b) within-team nHD.

Intuitively, one would think that agents learning by diffusion of knowledge would become more similar over time, and this does occur in initially scattered populations (who start at maximum diversity). However, in [27] we reported that the within-population similarity of clustered populations actually decreases through team learning, even as fitness continues to improve. To understand this seemingly counter-intuitive finding, we took a closer look at how within-team similarity (Fig. 3.6) and within-population similarity (Fig. 3.7) change during the learning process, for a single clustered initial population searching a simple landscape (no feature interactions), using homophilous team formation after each of 500 trial steps, with varying degrees of noise in the fitness function.

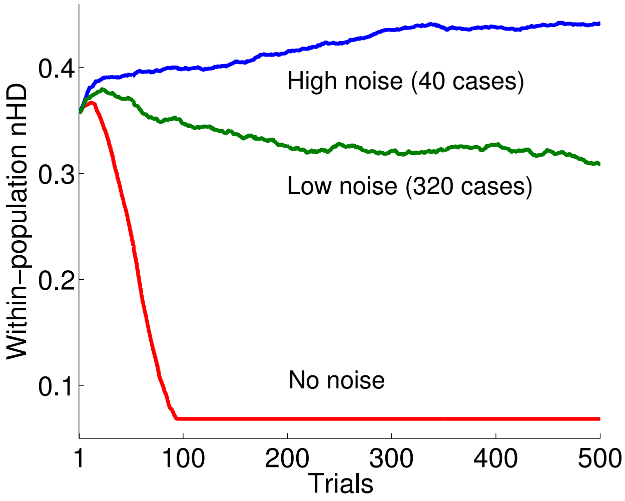


(a)

(b)

(c)

Figure 3.6: Within-team nHD for each of the ten teams (colored lines) over 500 trial steps starting from the same populations shown in Fig. 3.7. Since teams are reformed homophilously after each trial step, each colored line does not necessarily represent the same set of ten agents in different trial steps. The level of noise in fitness evaluation varies between the three panels: a) no noise, b) low noise, with 320 patients per trial, and c) high noise, with only 40 patients per trial.



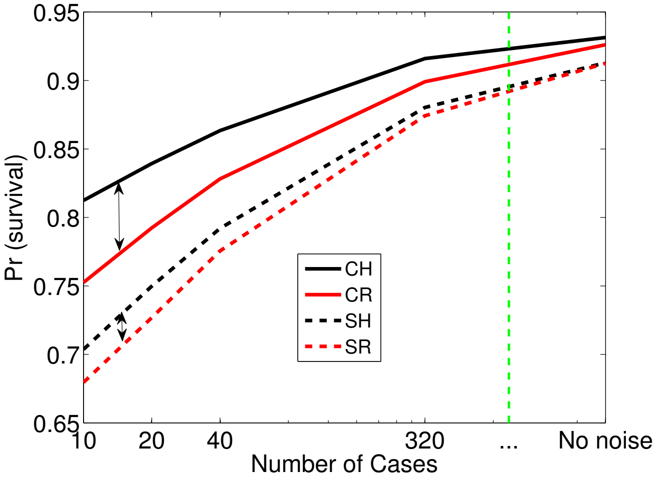
(a)

Figure 3.7: Within-population nHD for one initially clustered population over 500 trial steps on a landscape with no feature interactions, with homophilous team reformation after each trial step. The level of noise in fitness evaluation varies between the three lines, as indicated.

When there is no noise, each of the ten teams converged to a single vector (Fig. 3.6a), so that subsequent homophilous team selection resulted in no switching of agents between teams and further learning ceased. Further examination showed that these ten teams actually converged on nine different but similar vectors (all with excellent, although not identical, fitnesses), accounting for the small non-zero within-population nHD in the noise-free case (Fig. 3.7).

As the level of noise increases, stochastic effects prevent convergence. At low noise, the within-population nHD initially increases and then slowly decreases (Fig. 3.7) because as learning progresses the most similar teams tend to stay together (Fig. 3.6b, lower lines) which offsets the fact that stochastic effects cause the less homophilous teams to experience more mixing, causing within-team nHDs to rise and then plateau (Fig. 3.6b, upper lines). However, at high noise levels the within-population nHD steadily increases (Fig. 3.7).

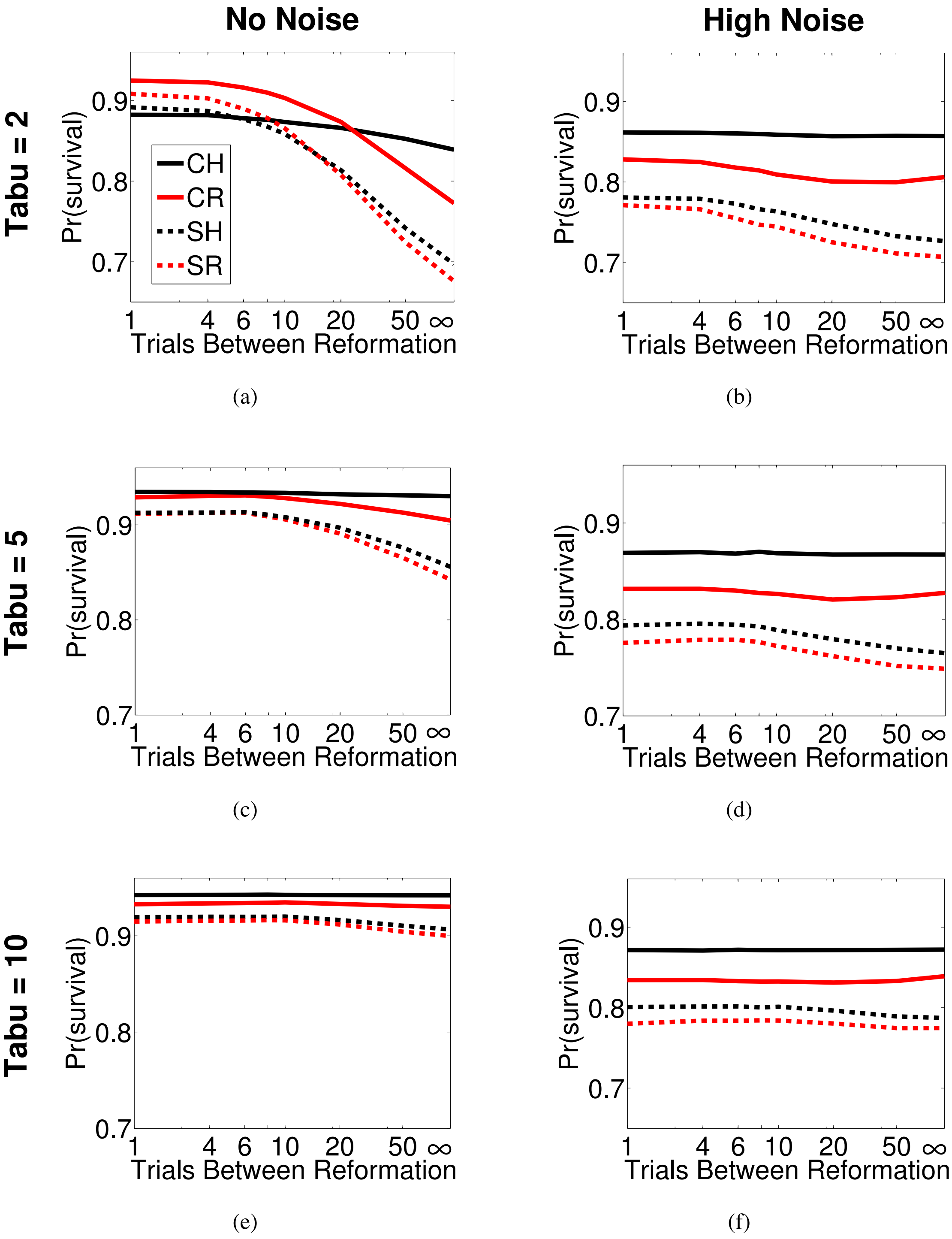
This occurs because stochastic effects introduce diversity that results in frequent switching of agents between teams, with a consequent rise in within-team nHDs (Fig. 3.6c), even as fitnesses increase through the learning process (recall Fig. 3.3a, CH line). The high fitnesses of all these diverse individuals is indicative of the fact that, even with no feature interactions, the variability in feature coefficients and the logistic compression of the landscape model (Eq. (1)) result in many excellent solutions, even though in this simple landscape there is only a single optimum.



(a)

Figure 3.8: Mean probability of patent survival (averaged over 100 trial steps and 100 random landscapes, each with 495 two-feature interactions), shown as a function of the number of patients in each trial. Note that increasing the number of cases decreases the noise in the fitness function. The fitness in the no noise case is computed using Eq. (1) directly rather than using Bernoulli trials, and therefore represents the asymptotic value for an infinite number of patients.

To further investigate the affects of noise in trials, we looked at the mean probability of survival (averaged over 100 agents and 100 random landscapes) as a function of the number of patients per trial (Fig. 3.8). Not surprisingly, learning became easier as noise decreased. What we found more interesting is that the advantage of homophilous teams over random



(a)

(b)

(c)

(d)

(e)

(f)

Figure 3.9: Mean Probability of survival on 100 random landscapes with 495 two-feature interactions as a function of the number of trial steps between team reformation. From top to bottom tabu is 2, 5 and 10, respectively. From left to right is noise-free or high noise (40 patients per trial), respectively.

teams was increasingly pronounced with higher noise (fewer patients per trial), especially when starting from clustered initial populations (Fig. 3.8, compare the magnitudes of the double black arrows).

To learn most effectively when there are feature interactions affecting fitness and there is noise in trial outcomes, hospitals may have to reevaluate previously tested features as their local contexts change through the act of learning. In [27], we allowed features to be reevaluated after waiting only one trial step ($tabu = 1$). Here, we examined $tabu \in \{2, 5, 10\}$ to test the sensitivity of performance improvement to the minimum number of trail steps each hospital was forced to wait before reevaluating the same feature on landscapes with intermediate complexity (495 two-feature interactions) (Fig. 3.9).

Our results show that, when fitness evaluation is noise-free, higher $tabu$ values result in a higher average probability of survival and reduced sensitivity to the frequency of team reformation (Fig. 3.9, left panels), because higher $tabu$ values force exploration of more features. When fitness is noisy, the same trends are qualitatively true but the sensitivity to $tabu$ is markedly reduced (Fig. 3.9, right panels). When there is no noise, a low $tabu$ value, and frequent reformation of teams, the homophilous teams of clustered agents (CH) are actually outperformed by the more diverse teams (Fig. 3.9a). Further exploration showed that this occurs because the CH teams become “locked in” and stagnate after about 50 trial steps as they continually retry features that look promising but aren’t, and the homophilous team reformation means that agents no longer switch teams and learning ceases. On the other hand, under these conditions the greater mixing due to random reformation actually promotes more exploration by preventing agents from continually retrying the same features, even though the $tabu$ is low and fitness is noise-free. When teams are never or rarely reformed and $tabu$ is low, however, the more diverse teams also stagnate and performance rapidly drops off, even lagging behind their counterparts with noisy fitness (compare rightmost values of Fig. 3.9a,b). In this case, the stochasticity introduced by noise actually promotes learning by preventing stagnation within teams.

Another important influence on team learning is the team size. In Fig. 3.10, we show the effects of partitioning the 100 agents into equal-size teams of a variety of sizes, starting from clustered initial populations on landscapes with 495 two-feature interactions. In general, the performance of agents was better for larger teams in these clustered populations, although the performance of homophilous teams does begin to drop very slightly for a single team of 100 agents (Fig. 3.10). For very small teams (teams of size 2 for homophilous teams, or teams up to size 4 for random teams), team search was actually outperformed by random search (Fig. 3.10), because there are too few teammates to learn from and exploration is therefore constrained. Homophilous teams consistently outperformed randomly formed teams and were less sensitive to team size.

Our finding that factors that increase within-team similarity promote robust team learning motivated us to try to answer the following three questions regarding real hospitals that are trying to improve clinical outcomes by working collaboratively to share information about clinical practices via VON-sponsored activities: (1) How clustered is the entire population of hospitals that comprise the VON network? (2) How clustered is the subpopulation of VON hospitals that actively participate in team learning through QICs? (3) Is the within-team similarity of VON QIC teams as high as possible, given the participating hospitals? Although the VON data shown in Fig. 3.1a provide encouraging preliminary evidence that the VON hospitals participating in QIC teams are clustered with regard to these clinical practices (question 2), the data on these 93 binarized clinical practices were only available for 51 hospitals that had completed a survey in 2003. Thus, we looked for fur-

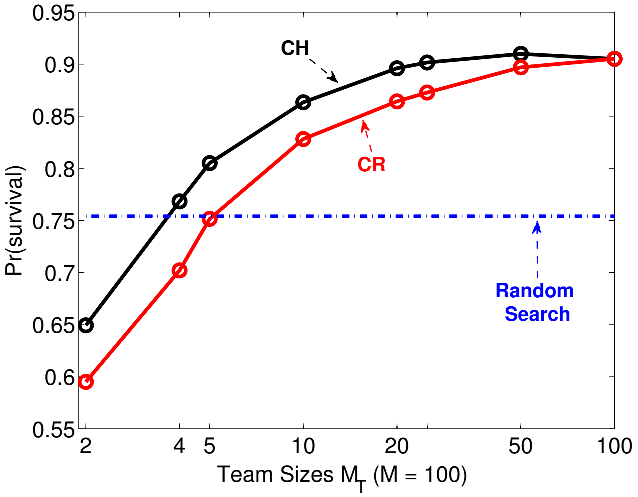


Figure 3.10: Mean fitnesses on 100 random landscapes with 495 two-feature interactions and clustered initial populations, over different team sizes and the amount of information team members have regarding their teammates' fitnesses. The horizontal line denotes the performance of random searchers.

ther evidence of clustering and within-team similarity in a larger data set of VON hospital information that includes more hospitals over a longer time span.

3.5 Data curation and analysis for Vermont Oxford Network hospitals

We report on two kinds of collaborations supported by the VON: (i) VON membership, which includes participation in annual meetings with seminars and posters on the effectiveness of various clinical practices, email listsserves, and access to a variety of shared information posted on the web, and (ii) neonatal intensive care QICs, referred to as NICQs. NICQs are extended collaborations among hospitals (meeting 2 times/year over 2-3 years), where practitioners from different hospitals are grouped into teams of 3-16 members and share results of case studies conducted at their home institutions. Based on this, different team members select what appear to be potentially better practices and try them out in their own institutions. Some hospitals chose to participate in multiple focus groups in the same NICQ, and a few joined later or dropped out early over the course of a given multi-year collaborative.

The VON has maintained extensive records regarding hospital member characteristics and participation in VON-related activities since its inception in 1990, including participation in NICQs. However, many of those records were only on paper, some were in disparate databases, there are many instances of missing data, and much of the data is confidential. For the purposes of this study, we curated and analyzed a subset of data reported by VON member hospitals in the time period of 1990 - 2010 and VON records of six multi-year NICQs (each comprising multiple teams), which were held in the following years: 1995-1998, 1999-2001, 2002-2004, 2005-2006, 2007-2008, and 2009-2010. We manually scanned the archives to identify which hospitals participated in which focus groups of these NICQs; we considered a hospital to be in a focus group if there was at least one healthcare practitioner from that hospital in that focus group. Limited hospital-level data on clinical practices was provided by VON for use in this study, and was de-identified to protect the confidentiality of patients and hospitals. The protocol for this research was submitted by the Committees on Human Research at the University of Vermont and determined to be exempt from formal Committee review and approval.

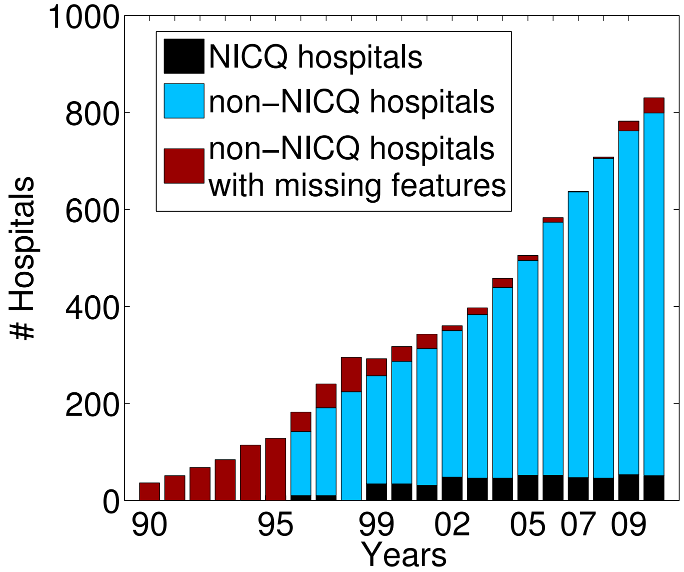
In collaboration with VON staff, we selected 20 of these clinical practices (see Table 3.3) that we thought might conceivably relate to various problems tackled by NICQ focus groups. Each of these practices are reported as the proportion of patients receiving the practice (or treatment), and hence are real-valued numbers between 0 and 1.

Table 3.3: Health practices used as features in our analysis of the VON data. All practices are reported as proportions of patients in each hospital that received those practices (treatments).

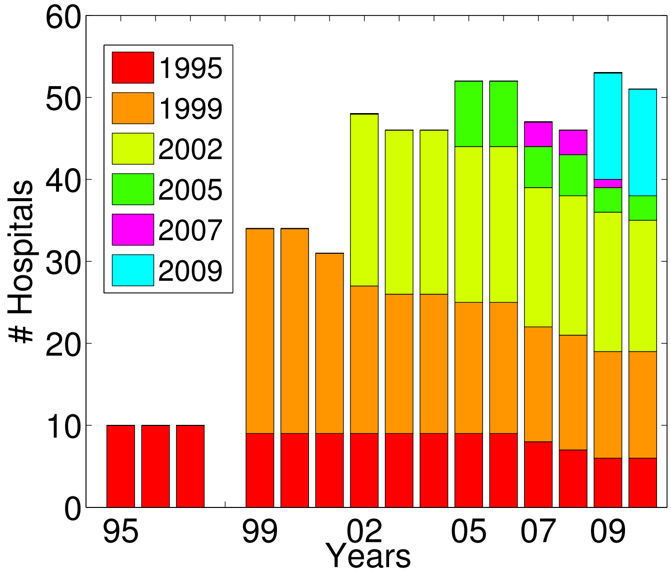
#	**Description**
1	Prenatal Care
2	Antenatal Steroids
3	Vaginal Delivery
4	Oxygen during initial resuscitation
5	Face mask ventilation during initial resuscitation
6	Endotracheal tube during initial resuscitation
7	Epinephrine during initial resuscitation
8	Cardiac compressions during initial resuscitation
9	Cranial imaging on/before day 28
10	Oxygen After Initial resuscitation
11	Conventional ventilation after initial resuscitation
12	High frequency ventilation after initial resuscitation
13	Nasal CPAP after initial resuscitation
14	Surfactant at any time
15	Steroids for Chronic Lang Disease
16	Indomethacin for any reason
17	Retinopathy of Prematurity Surgery
18	Necrotizing Enterocolitis Surgery
19	Other Minor Surgery
20	Any Major Surgery

The number of hospitals in the VON grew from about 50 hospitals in 1990 to more than 800 hospitals in 2010 (see Fig. 3.11a). Starting in 1995, a small subset of these participated in NICQs (Fig. 3.11a, black). Of the remainder, some had missing data (Fig. 3.11a, dark red) and were thus excluded from this study. A more detailed histogram of hospitals that participated in NICQs is shown in Fig. 3.11b, where each color indicates the year that a given hospital first joined a NICQ focus group. Over this period, the NICQ subpopulation grew from 10 to over 50 hospitals, with a relatively low dropout rate, as evidenced by the roughly parallel bands of color in Fig. 3.11b.

Some important differences between the real NICQ teams and the teams in our ABM are that, in the real NICQs, different teams are studying different topics impacting a variety of clinical outcomes (so there is no single health outcome available to measure the impacts of team learning), team sizes vary, the set of hospitals participating in NICQ teams changes over time, and we only have data on real-valued rates of certain clinical practices that are routinely collected by the VON. Furthermore, there is a wide degree of variation in patient demographics and other unchangeable characteristics among VON hospitals that impact clinical outcomes. These sorts of complications have made it difficult for researchers to find direct evidence that QICs have directly improved health outcomes [84, 119]. Nonetheless, we can use the VON data to assess clustering among the clinical practices for which we have information. To assess the distance between two hospitals in a given year, we computed normalized Euclidean distances (nEDs) between the real-valued practice rates reported for the 20 practices shown in Table 3.3, normalized by dividing by the maximum possible Euclidean distance between 20 practices (4.47). Within-population nED is defined as the mean of all pairwise nEDs in a population (or specified subpopulation) of hospitals in a given year, and within-team nED is defined as the mean of the mean of all pairwise nEDs within each focus group in a given year. Using these measures, a uniformly scattered population will have a within-population nED of about 0.4. Thus, in the following results, nED ranges from 0 (maximum similarity) to 0.4 (maximum diversity).



(a)

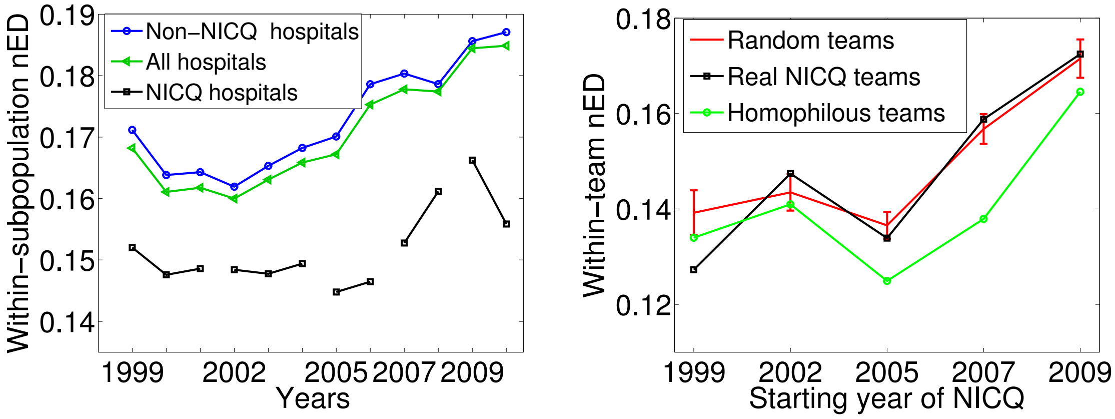


(b)

Figure 3.11: a) Number of the hospitals in the VON network from 1990 till 2010 that either participated in NICQ collaboratives (dark blue), have complete records but didn't participate in NICQs (light blue), or didn't participate in NICQs and have incomplete records (dark red). b) More detailed view of the number of hospitals in NICQ collaboratives from 1995 till 2010. Each color represents the years that each given hospital first joined a NICQ focus group. X-axis labels only show the starting years of the six multi-year NICQ collaboratives.

3.6 Results for hospitals in VON

Within-population nEDs for the entire VON network, as well as for the subpopulations of NICQ hospitals and non-NICQ hospitals, are shown in Fig. 3.12a, for the 5 NICQs starting in 1999 through 2009 (with connected dots denoting years of a given NICQ). These results show that the hospitals in the VON are quite clustered with respect to these 20 practices (all values are less than half the maximum possible nED of 0.4), and that those who chose to participate in NICQ collaboratives are even more clustered than the rest of the VON.



(a)

(b)

Figure 3.12: a) Mean pairwise within subpopulation normalized Euclidean Distances (nEDs) (i.e., subpopulation closeness) in the NICQ and non-NICQ subpopulations, where nEDs are calculated for either 20 practices or 15 outcomes. b) Average of the mean pairwise normalized within team EDs for either randomly formed teams, real NICQ teams or homophilous teams (picked by $PickSimilar$ algorithm). Euclidean Distances between individuals are calculated for either 20 practices or 15 outcomes. X-axis labels only show the starting years of 5 NICQ collaboratives in 1999-2010.

With the exception of the 1999 NICQ, the within-team nEDs of NICQ hospitals (Fig. 3.12b, black lines) were within one standard deviation of the nEDs of 100 randomly formed teams of the same sizes drawn from the real NICQ subpopulation in each year

(Fig. 3.12b, red lines, with error bars indicating ± one standard deviation) and were more diverse than homophilous teams selected from the same NICQ subpopulation using the $PickSimilarTeams$ algorithm (see Fig. 3.12b, green lines), indicating that even greater within-team similarity is possible.

There is also a small but statistically significant increase in both the within-population ($p < 0.001$) and within-team ($p < 0.05$) diversity over the years studied (Fig. 3.12). The fact that the network is growing over these years (Fig. 3.11a) undoubtedly contributes to these increases in diversity, so it is not possible to ascertain whether any of this is attributable to the collaborative learning processes themselves, as occurred when there was noise in the fitness evaluation in the ABM.

3.7 Discussion

The aim of this study was to try to gain insight into which factors enhance team learning in environments where the goal is knowledge diffusion, rather than knowledge creation. We used an ABM to examine the sensitivity of quality improvement at individual simulated hospitals to different team collaboration scenarios. The results of the ABM show that learning in teams through collaborative diffusion of knowledge is most effective, and most robust to a variety of external influences, when within-team similarity is high. This contrasts with previous findings that diverse teams improve collaborative problem-solving [41], [39].

We examined several factors that contribute to within-team similarity, most notably the degree of clustering in feature space of the initial population and the type of team formation strategy. A homophilous team formation strategy continually ensures that within-team similarity remains as high as possible, even when other factors exert pressure in the opposite direction. Because of this, homophilous team formation has several advantages for team learning through diffusion of knowledge, especially when agents are clustered in complex landscapes and fitness evaluation is noisy, as is likely to be the case in healthcare institutions. Under a wide range of scenarios studied, homophilous team formation in clustered populations was the top performer, with the exception of a single unrealistic scenario (Fig. 3.9a, with noise free fitness). Furthermore, the performance of homophilous teams proved to be less sensitive to a variety of factors, including the complexity of the landscape, the level of noise, the size of the team, the frequency of team reformation, and the $tabu$ time before agents were allowed to reevaluate a feature. The consistent nature of this finding suggests that homophilous teams may be beneficial in real world collaborative learning environments, like healthcare QICs, where the emphasis is on knowledge diffusion (rather than knowledge creation).

Despite long-standing recognition of the existence of widespread variations in clinical practices ([136], [77], [137]), we previously [27] postulated that real hospitals would exhibit a high degree of clustering in this landscape of clinical practices. In this work, examination of a snapshot of 93 binarized clinical practices in 51 real hospitals participating in QICs confirmed that they were highly clustered with respect to these practices. An analysis of a larger data set of 20 different real-valued practices, over 11 years in a growing network of collaborating hospitals ultimately comprising more than 800 hospitals, also revealed a high degree of clustering, with those hospitals actively participating in team learning through QICs even more tightly clustered than the population at large.

While our examination of within-team similarity in real QIC teams did not show evidence of homophilous team formation, the high degree of clustering within the clinical practices studied implies that even randomly formed teams from this population of hospitals will have a high degree of within-team similarity, with respect to these practices.

Currently, real VON QIC teams (focus groups) are largely self-organized with respect to interest in exploring the various topical areas, although in some cases VON staff do split up large teams or otherwise influence team membership. Since there does appear to be room for slightly greater within-team similarity in VON focus groups, it may be advisable for VON staff to actively encourage more similar hospitals to group together, especially with respect to externally controlled features that are likely to influence local contexts of care and patient outcomes, such as patient demographics, hospital size, and geographical cultures. This may become increasingly important as more hospitals join the VON and elect to participate in QICs, resulting in an increasingly diverse population.

In the healthcare domain, detailed data on clinical practices and patient outcomes is already collected and maintained securely by organizations such as the VON, but this information is not shared publicly. However, unless teammates have detailed knowledge about the clinical practices and fitnesses of their teammates, the feature selection mechanism used in team search will essentially degenerate to random search. Our simulation results also showed that larger teams of already clustered agents performed better than smaller teams, since more information was available to learn from. These results suggest that, in an ideal world, one would have similar hospitals collaborate in large teams and have open access to all data about each other, in order to derive optimal benefits from the collaboration. However in the real world, the maximum number of individuals in QIC teams is limited both by organizational costs related to team assembly into a collaborative environment and by the number of individuals that can effectively work together in that environment, and real hospitals have significant privacy concerns regarding sharing detailed data on practices and outcomes. Furthermore, hospitals who already have excellent health outcomes of a particular type are less likely to join a focus group that is studying ways to improve that health outcome, potentially limiting the maximum fitness within a given team with respect to their primary outcome of interest.

Thus, our inferred optimal learning strategy is in conflict with the realities of team learning in real healthcare QICs in a variety of ways. One possible way to mitigate these conflicts would be through a Virtual Collaboration System (VCS) that would allow hospitals to efficiently identify potentially better practices in use at other institutions similar to theirs, without any hospitals having to sacrifice the privacy of their own institutional data. Suppose that a given hospital queries such a VCS for possible ways to improve its performance with regard to a specific type of health outcome. The VCS could compare the clinical practices and other characteristics of the hospital to those in the database to identify a large virtual team of similar institutions, compare the specific health outcome of interest to identify which virtual team members are better or worse performers with regard to that outcome, and could then make intelligent customized recommendations of potentially better practices to the hospital, using an algorithm similar to the feature selection algorithm described in Section 3.3.4. Hospitals identified as successfully employing a clinical practice that may be beneficial to another institution could be confidentially contacted to see if they would be willing to host a visit from the inquiring hospital to share more about the details of this practice. However, since such a VCS would not require physical assembly of actual teams, it would reduce the time and other costs associated with collaborative learning, relative to current healthcare QICs. Hospitals would be required to share detailed information on their practices and outcomes to be able to use the VCS, but would be incentivized to do so by being able to benefit from the collective knowledge. In fact, many healthcare organizations are already providing similar confidential data to organizations like the VON for internal analysis, so with the infrastructure already in place extending this with a recommendation system would not be particularly onerous.

Our findings may also prove useful in other application domains, such as in collaborations designed to share best practices within franchises of a business, each with slightly different local contexts. In addition, with the growing availability of genomic data and electronic medical records, there has been increasing interest in the potential for personalized medicine [58, 115, 133]. It is conceivable that large databases of human DNA sequences and other relevant patient-specific attributes, health conditions, treatments, and outcomes, could be queried using an approach similar to that proposed here for virtual QICs, to suggest promising personalized treatments.

Finally, we also believe that these findings may provide useful guidance in designing effective evolutionary algorithms to solve combinatorial optimization problems with complex and/or noisy fitness landscapes. In this context, one can view team learning as a form of smart crossover. In this more abstract problem solving domain, initial population distributions and other factors are not constrained by reality (as they are in the healthcare domain). In future work, we plan to compare team learning strategies and clustered initial populations to genetic algorithms starting from scattered initial populations and using more standard forms of crossover on combinatorial optimization problems of varying difficulty.

3.8 Summary and Conclusions

Healthcare institutions are increasingly participating in quality improvement collaboratives (QICs). In these collaborations multi-institutional teams share information, and representatives of each institution identify potentially better practices that are subsequently evaluated in the local contexts of their home institutions. In this paper we modeled this collaborative learning approach using an agent-based model (ABM) to study how different team characteristics affect quality improvement of agents (simulated hospitals) in clinical fitness landscapes with varying degrees of complexity (interactions between clinical practices) and noise (based on the number of patients in each trial).

We first analyzed a set of binarized clinical practices in real hospitals that participated in QICs and found that these hospitals are clustered with respect to these practices. Guided by the real data, we introduced a new method for generating synthetic agents that are similarly clustered in feature space. We also introduced a new method for selecting teams of homophilous agents. These methods were incorporated into the ABM and a variety of sensitivity studies were performed.

Our simulations show that, in this type of learning environment (where the goal is diffusion of knowledge to improve outcomes of individual agents rather than joint-problem solving), teams with higher within-team similarity are able to improve performance more quickly than diverse teams, are less sensitive to a variety of factors, and larger teams of similar agents generally perform better than smaller teams. Notably, the advantage of within-team similarity increases with the complexity of the fitness landscape and with the level of noise in fitness evaluation. Interesting interactions are shown to occur between the frequency of team reformation, the minimum number of trials steps before which an agent can retry the same feature, the team formation strategy, the complexity of the landscape, and the level of noise.

Further analysis of a larger data set of 20 real-valued practices over 11 years in the growing Vermont Oxford Network (VON) of hospitals provided further evidence that (a) hospitals in the VON are clustered in the landscape of clinical practices, (b) the set of VON hospitals that actively participate in team learning through QICs are even more clustered than the population of hospitals at large, resulting in high within-team similarity, and that (c) there is room for even greater within-team similarity in VON QICs if teams are encouraged to form using the principle of homophily.

Based on these results, we propose a new virtual collaboration framework that could allow hospitals to efficiently improve quality by learning from a secure and confidential knowledge base using an intelligent recommendation system to select which features to test next in their own institutions. While this work was specifically motivated to inform quality improvement in healthcare institutions, our results may also have bearing on other types of learning environments where the aim is the diffusion of contextually relevant knowledge in complex environments, including in personalized medicine, spreading of best practices within franchises of a business, or evolutionary computational approaches to combinatorial optimization problems.

Acknowledgments

The authors would like to thank Mike Kenny, Kathy Leahy, and the Vermont Oxford Network for their time and effort in providing the hospital data, and Shreya Mukherjee for her help in curating the data. We also thank Stuart A. Kauffman, Jeffrey S. Buzas, and Donna M. Rizzo for stimulating discussions that benefitted this work. This work was funded in part by the NIH Eunice Kennedy Shriver National Institute of Child Health and Human Development award 1R21HD068296.

Bibliography

[1] Roberta Annicchiarico, Ulises Cortés, and Cristina Urdiales. *Agent Technology and E-health*. Springer, 2008.

[2] Ignacio Arnaldo, Iván Contreras, David Millán-Ruiz, J Ignacio Hidalgo, and Natalio Krasnogor. Matching island topologies to problem structure in parallel evolutionary algorithms. *Soft Computing*, pages 1–17, 2013.

[3] Lea R Ayers, Suzanne C Beyea, Marjorie M Godfrey, Doreen C Harper, Eugene C Nelson, and Paul B Batalden. Quality improvement learning collaboratives. *Quality Management in Healthcare*, 14(4):234–247, 2005.

[4] Rosa R Baier, David R Gifford, Gail Patry, Sara M Banks, Therese Rochon, Debra DeSilva, and Joan M Teno. Ameliorating pain in nursing homes: a collaborative quality-improvement project. *Journal of the American Geriatrics Society*, 52(12):1988–1995, 2004.

[5] David W Baker, Steven M Asch, Joan W Keesey, Julie A Brown, Kitty S Chan, Geoffrey Joyce, and Emmett B Keeler. Differences in education, knowledge, self-management activities, and health outcomes for patients with heart failure cared for under the chronic disease model: the improving chronic illness care evaluation. *Journal of cardiac failure*, 11(6):405–413, 2005.

[6] A.S. Banks. Cross-national time-series data archive (cnts) 1815-2007. *Databanks International, Jerusalem, Israel*, 2008.

[7] Alberto Barceló, Elizabeth Cafiero, Melanie de Boer, Alejandro Escobar Mesa, Marcelina García Lopez, Rosa Aurora Jiménez, Agustín Lara Esqueda, José Antonio Martinez, Esperanza Medina Holguin, Micheline Meiners, et al. Using collaborative learning to improve diabetes care and outcomes: The vida project. *Primary care diabetes*, 4(3):145–153, 2010.

[8] Paul B Batalden and Frank Davidoff. What is quality improvement and how can it transform healthcare? *Quality and Safety in Health Care*, 16(1):2–3, 2007.

[9] Jordana T Bell, Nicholas J Timpson, N William Rayner, Eleftheria Zeggini, Timothy M Frayling, Andrew T Hattersley, Andrew P Morris, and Mark I McCarthy. Genome-wide association scan allowing for epistasis in type 2 diabetes. *Annals of human genetics*, 75(1):10–19, 2011.

[10] Rob Benedetti, Barb Flock, Steve Pedersen, et al. Improved clinical outcomes for fee-for-service physician practices participating in a diabetes care collaborative. *Joint Commission Journal on Quality and Patient Safety*, 30(4):187–194, 2004.

[11] I.M. Bernstein, J.D. Horbar, G.J. Badger, A. Ohlsson, A. Golan, et al. Morbidity and mortality among very-low-birth-weight neonates with intrauterine growth restriction. *American journal of obstetrics and gynecology*, 182(1):198–206, 2000.

[12] DM Berwick. Broadening the view of evidence-based medicine. *Quality and Safety in Health Care*, 14(5):315–316, 2005.

[13] Sebastian Bonhoeffer, Colombe Chappey, Neil T Parkin, Jeanette M Whitcomb, and Christos J Petropoulos. Evidence for positive epistasis in hiv-1. *Science*, 306(5701):1547–1550, 2004.

[14] Penny Bundy. Using drama in the counselling process: the moving on project. *Research in drama education*, 11(1):7–18, 2006.

[15] Jeffrey Buzas and Jeffrey Dinitz. An analysis of NK landscapes: Interaction structure, statistical properties and expected number of local optima. *IEEE Transactions on Evolutionary Computation*, in press, DOI10.1109/TEVC.2013.2286352, 2014.

[16] Pilar Caamaño, Abraham Prieto, José Antonio Becerra, Francisco Bellas, and Richard J Duro. Real-valued multimodal fitness landscape characterization for evolution. In *Neural Information Processing. Theory and Algorithms*, pages 567–574. Springer, 2010.

[17] P.R. Cohen and H.J. Levesque. Teamwork. *Nous*, pages 487–512, 1991.

[18] A. Dechartres, I. Boutron, L. Trinquart, et al. Single-center trials show larger treatment effects than multicenter trials: Evidence from a meta-epidemiologic study. *Annals of internal medicine*, 155(1):39, 2011.

[19] D. DeHaas, J. Craig, C. Rickert, P. Haake, K. Stor, and M.J. Eppstein. Feature selection and classification in noisy epistatic problems using a hybrid evolutionary approach. *poster and published extended abstract accepted for Genetic and Evolutionary Computation Conference (GECCO)*, 2007.

[20] OM Dekkers, Erik von Elm, Ale Algra, JA Romijn, and JP Vandenbroucke. How to assess the external validity of therapeutic trials: a conceptual approach. *International journal of epidemiology*, 39(1):89–94, 2010.

[21] J. Denrell and C. Liu. Top performers are not the most impressive when extreme performance indicates unreliability. *Proceedings of the National Academy of Sciences*, 109(24):9331–9336, 2012.

[22] Hui Ding, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, and Eamonn Keogh. Querying and mining of time series data: experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment*, 1(2):1542–1552, 2008.

[23] Lori Ebert, Lisa Amaya-Jackson, Jan M Markiewicz, Cassandra Kisiel, and John A Fairbank. Use of the breakthrough series collaborative to support broad and sustained use of evidence-based trauma treatment for children in community practice settings. *Administration and Policy in Mental Health and Mental Health Services Research*, 39(3):187–199, 2012.

[24] Judith A Effken. Different lenses, improved outcomes: a new approach to the analysis and design of healthcare information systems. *International journal of medical informatics*, 65(1):59–74, 2002.

[25] Margaret J Eppstein and Paul Haake. Very large scale relieff for genome-wide association analysis. In *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 112–119, 2008.

[26] M.J. Eppstein and P.D.H. Hines. A random chemistry; algorithm for identifying collections of multiple contingencies that initiate cascading failure. *IEEE Transactions on Power Systems*, 27(3):1698–1705, 2012.

[27] M.J. Eppstein, J.D. Horbar, J.S. Buzas, and S.A. Kauffman. Searching the clinical fitness landscape. *PLoS ONE*, 7(11):e49901, 2012.

[28] Philippe Esling and Carlos Agon. Time-series data mining. *ACM Computing Surveys (CSUR)*, 45(1):12, 2012.

[29] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.

[30] Usama M Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy. Advances in knowledge discovery and data mining. 1996.

[31] Stephanie Forrest and Melanie Mitchell. The performance of genetic algorithms on walsh polynomials: Some anomalous results and their explanation. In *Proceedings of the 4th International Cinference on Genetic Alogarithms*, pages 182–189. San Mateo, CA: Morgan Kaufmann, 1991.

[32] Sarah W Fraser and Trisha Greenhalgh. Coping with complexity: educating for capability. *BMJ*, 323(7316):799–803, 2001.

[33] Mingxin Gan and Rui Jiang. Constructing a user similarity network to remove adverse influence of popular objects for personalized recommendation. *Expert Systems with Applications*, 2013.

[34] Mingxin Gan and Rui Jiang. Improving accuracy and diversity of personalized recommendation through power law adjustments of user similarities. *Decision Support Systems*, 2013.

[35] Yong Gao and Joseph C. Culberson. An analysis of phase transition in NK landscapes. *Journal of Artificial Intelligence Research*, 17(1):309–332, 2002.

[36] Ilaria Giannoccaro. Complex systems methodologies for behavioural research in operations management: NK fitness landscape. In *Behavioral Issues in Operations Management*, pages 23–47. Springer, 2013.

[37] R.J. Gray. A bayesian analysis of institutional effects in a multicenter cancer clinical trial. *Biometrics*, pages 244–253, 1994.

[38] Jon W Gregersen, Kamil R Kranc, Xiayi Ke, Pia Svendsen, Lars S Madsen, Allan Randrup Thomsen, Lon R Cardon, John I Bell, and Lars Fugger. Functional epistasis on a common mhc haplotype associated with multiple sclerosis. *Nature*, 443(7111):574–577, 2006.

[39] R. Guimera, B. Uzzi, J. Spiro, and L.A.N. Amaral. Team assembly mechanisms determine collaboration network structure and team performance. *Science*, 308(5722):697–702, 2005.

[40] R. Hecht-Nielsen. Counterpropagation networks. *Applied optics*, 26(23):4979–4983, 1987.

[41] L. Hong and S.E. Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the United States of America*, 101(46):16385–16389, 2004.

[42] J.D. Horbar. The vermont oxford network: evidence-based quality improvement for neonatology. *Pediatrics*, 103(Supplement):350, 1999.

[43] J.D. Horbar, G.J. Badger, J.H. Carpenter, A.A. Fanaroff, S. Kilpatrick, M. LaCorte, R. Phibbs, R.F. Soll, et al. Trends in mortality and morbidity for very low birth weight infants, 1991–1999. *Pediatrics*, 110(1):143, 2002.

[44] J.D. Horbar, G.J. Badger, E.M. Lewit, J. Rogowski, P.H. Shiono, et al. Hospital and patient characteristics associated with variation in 28-day mortality rates for very low birth weight infants. *Pediatrics*, 99(2):149, 1997.

[45] J.D. Horbar et al. The vermont-oxford neonatal network: integrating research and clinical practice to improve the quality of medical care. In *Seminars in perinatology*, volume 19, page 124, 1995.

[46] J.D. Horbar and J.F. Lucey. Evaluation of neonatal intensive care technologies. *The Future of Children*, pages 139–161, 1995.

[47] J.D. Horbar, P.E. Plsek, and K. Leahy. Nic/q 2000: establishing habits for improvement in neonatal intensive care units. *Pediatrics*, 111(Supplement):e397, 2003.

[48] J.D. Horbar, P.E. Plsek, J.A. Schriefer, and K. Leahy. Evidence-based quality improvement in neonatal and perinatal medicine: the neonatal intensive care quality improvement collaborative experience. *Pediatrics*, 118(Supplement):S57, 2006.

[49] J.D. Horbar, J. Rogowski, P.E. Plsek, P. Delmore, W.H. Edwards, J. Hocker, A.D. Kantak, P. Lewallen, W. Lewis, E. Lewit, et al. Collaborative quality improvement for neonatal intensive care. *Pediatrics*, 107(1):14–22, 2001.

[50] J.D. Horbar, J. Rogowski, P.E. Plsek, P. Delmore, W.H. Edwards, J. Hocker, A.D. Kantak, P. Lewallen, W. Lewis, E. Lewit, et al. Collaborative quality improvement for neonatal intensive care. *Pediatrics*, 107(1):14–22, 2001.

[51] J.D. Horbar, R.F. Soll, and W.H. Edwards. The vermont oxford network: a community of practice. *Clin Perinatol*, 37(1):29–47, 2010.

[52] J.D. Horbar, R.F. Soll, and W.H. Edwards. The Vermont Oxford Network: A community of practice. *Clinics in perinatology*, 37(1):29, 2010.

[53] Wim Hordijk. Correlation analysis of coupled fitness landscapes. In *Recent Advances in the Theory and Application of Fitness Landscapes*, pages 369–393. Springer, 2014.

[54] R.I. Horwitz, B.H. Singer, R.W. Makuch, and C.M. Viscoli. Can treatment that is helpful on average be harmful to some patients? A study of the conflicting information needs of clinical inquiry and drug regulation. *Journal of Clinical Epidemiology*, 49(4):395–400, 1996.

[55] T.K. Jenssen, W. Kuo, T. Stokke, and E. Hovig. Associations between gene expressions in breast cancer and patient survival. *Human genetics*, 111(4):411–420, 2002.

[56] K. Johnell and I. Klarin. The relationship between number of drugs and potential drug-drug interactions in the elderly: A study of over 600000 elderly patients from the swedish prescribed drug register. *Drug Safety*, 30(10):911–918, 2007.

[57] Terry Jones and Stephanie Forrest. Fitness distance correlation as a measure of problem difficulty for genetic algorithms. In *ICGA*, volume 95, pages 184–192. Citeseer, 1995.

[58] E.T. Juengst, R.A. Settersten, J.R. Fishman, and M.L. McGowan. After the revolution? Ethical and social challenges in personalized genomic medicine. *Personalized Medicine*, 9(4):429–439, 2012.

[59] Leila Kallel, Bart Naudts, and Colin R Reeves. Properties of fitness functions and search landscapes. In *Theoretical aspects of evolutionary computing*, pages 175–206. Springer, 2001.

[60] Stuart Kauffman. *The origins of order: Self organization and selection in evolution*. Oxford University Press, 1993.

[61] Stuart Kauffman and Simon Levin. Towards a general theory of adaptive walks on rugged landscapes. *Journal of theoretical Biology*, 128(1):11–45, 1987.

[62] Stuart A Kauffman and Edward D Weinberger. The NK model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of theoretical biology*, 141(2):211–245, 1989.

[63] Eamonn Keogh and Shruti Kasetty. On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and knowledge discovery*, 7(4):349–371, 2003.

[64] Charles M Kilo. A framework for collaborative improvement: lessons from the institute for healthcare improvement's breakthrough series. *Quality Management in Healthcare*, 6(4):1–14, 1998.

[65] Bobby PC Koeleman, Benedicte Alexandre Lie, Dag Erik Undlien, Frank Dudbridge, Erik Thorsby, Rindert RP De Vries, Francesco Cucca, Bart O Roep, MJ Giphart, and John A Todd. Genotype effects and epistasis in type 1 diabetes and hla-dq trans dimer associations with disease. *Genes and immunity*, 5(5):381–388, 2004.

[66] J.S. Krupa, S. Chatterjee, E. Eldridge, D.M. Rizzo, and M.J. Eppstein. Evolutionary exploratory association discovery: A plug-in hybrid vehicle adoption application. *Submitted to the 21st International GECCO Confference*, 2012.

[67] J.A. LePine, R.F. Piccolo, C.L. Jackson, J.E. Mathieu, and J.R. Saul. A meta-analysis of teamwork processes: Tests of a multidimensional model and relationships with team effectiveness criteria. *Personnel Psychology*, 61(2):273–307, 2008.

[68] Rui Li, Michael TM Emmerich, Jeroen Eggermont, Ernst GP Bovenkamp, Thomas Bäck, Jouke Dijkstra, and Johan HC Reiber. Mixed-integer nk landscapes. In *Parallel Problem Solving from Nature-PPSN IX*, pages 42–51. Springer, 2006.

[69] Rung Tzuo Liaw and Chuan Kang Ting. Effect of model complexity for estimation of distribution algorithm in NK landscapes. In *2013 IEEE Symposium on Foundations of Computational Intelligence (FOCI)*, pages 76–83. IEEE, 2013.

[70] H. Liu, J. Li, and L. Wong. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Informatics Series*, pages 51–60, 2002.

[71] Rita Mangione-Smith, Matthias Schonlau, Kitty S Chan, Joan Keesey, Mayde Rosen, Thomas A Louis, and Emmett Keeler. Measuring the effectiveness of a collaborative for quality improvement in pediatric asthma care: does implementing the chronic care model improve processes and outcomes of care? *Ambulatory Pediatrics*, 5(2):75–82, 2005.

[72] T. Manser. Teamwork and patient safety in dynamic domains of healthcare: A review of the literature. *Acta Anaesthesiologica Scandinavica*, 53(2):143–151, 2008.

[73] Narine Manukyan, Margaret J Eppstein, and Jeffrey D Horbar. Team learning for healthcare quality improvement. *IEEE Access*, 1:545–557, 2013.

[74] Narine Manukyan, Margaret J Eppstein, and Donna M Rizzo. Data-driven cluster reinforcement and visualization in sparsely-matched self-organizing maps. *Neural Networks and Learning Systems, IEEE Transactions on*, 23(5):846–852, 2012.

[75] M.A. Marks, J.E. Mathieu, and S.J. Zaccaro. A temporally based framework and taxonomy of team processes. *Academy of Management Review*, pages 356–376, 2001.

[76] J.A. Martin, K.D. Kochanek, D.M. Strobino, B. Guyer, and M.F. MacDorman. Annual summary of vital statistics—2003. *Pediatrics*, 115(3):619, 2005.

[77] Klim McPherson, John E Wennberg, Ole B Hovind, Peter Clifford, et al. Small-area variations in the use of common surgical procedures: An international comparison of New England, England, and Norway. *The New England journal of medicine*, 307(21):1310, 1982.

[78] Brian S Mittman. Creating the evidence base for quality improvement collaboratives. *Annals of internal medicine*, 140(11):897–901, 2004.

[79] Naoki Miyagawa, Hiroshi Teramoto, Chun-Biu Li, and Tamiki Komatsuzaki. Decomposability of multivariate interactions. *Complex Systems*, 20(2):165, 2011.

[80] Douglas C Montgomery, Douglas C Montgomery, and Douglas C Montgomery. *Design and analysis of experiments*, volume 7. Wiley New York, 1984.

[81] Jason H Moore. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human heredity*, 56(1-3):73–82, 2003.

[82] Alberto Moraglio and Julian Togelius. Geometric differential evolution. In *Proceedings of the 11th Annual conference on Genetic and evolutionary computation*, pages 1705–1712. ACM, 2009.

[83] L.S. Morales, D. Staiger, J.D. Horbar, J. Carpenter, M. Kenny, J. Geppert, and J. Rogowski. Mortality among very low birthweight infants in hospitals serving minority populations. *American journal of public health*, 95(12):2206, 2005.

[84] Erum Nadeem, S Serene Olin, Laura Campbell Hill, Kimberly Eaton Hoagwood, and Sarah McCue Horwitz. Understanding the components of quality improvement collaboratives: A systematic literature review. *Milbank Quarterly*, 91(2):354–394, 2013.

[85] P.J. Newton, EJ Halcomb, PM Davidson, and A.R. Denniss. Barriers and facilitators to the implementation of the collaborative method: Reflections from a single site. *Quality and Safety in Health Care*, 16(6):409–414, 2007.

[86] I.S. Oh, J.S. Lee, and B.R. Moon. Hybrid genetic algorithms for feature selection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(11):1424–1437, 2004.

[87] World Health Organization et al. Who patient safety curriculum guide for medical schools. 2009.

[88] J. Øvretveit, P. Bate, P. Cleary, S. Cretin, D. Gustafson, K. McInnes, H. McLeod, T. Molfenter, P. Plsek, G. Robert, et al. Quality collaboratives: lessons from research. *Quality and safety in health care*, 11(4):345–351, 2002.

[89] J Øvretveit, Paul Bate, Paul Cleary, Shan Cretin, D Gustafson, K McInnes, H McLeod, T Molfenter, P Plsek, Glenn Robert, et al. Quality collaboratives: lessons from research. *Quality and safety in health care*, 11(4):345–351, 2002.

[90] Ray Pawson and Nick Tilley. *Realistic evaluation*. Sage, 1997.

[91] N.R. Payne, M.J. Finkelstein, M. Liu, J.W. Kaempf, P.J. Sharek, and S. Olsen. Nicu practices and outcomes associated with 9 years of quality improvement collaboratives. *Pediatrics*, 125(3):437–446, 2010.

[92] Martin Pelikan. Analysis of estimation of distribution algorithms and genetic algorithms on NK landscapes. In *Proceedings of the 10th annual conference on Genetic and evolutionary computation*, pages 1033–1040. ACM, 2008.

[93] F. Pernkopf and P. O'Leary. Feature selection for classification using genetic algorithms with a novel encoding. In *Computer Analysis of Images and Patterns*, pages 161–168. Springer, 2001.

[94] Charles Perrow. *Normal Accidents: Living with High Risk Technologies (Updated)*. Princeton University Press, 2011.

[95] Paul E Plsek. Collaborating across organizational boundaries to improve the quality of care. *American journal of infection control*, 25(2):85–95, 1997.

[96] Paul E Plsek and Trisha Greenhalgh. The challenge of complexity in health care. *Bmj*, 323(7313):625–628, 2001.

[97] Paul E Plsek and Tim Wilson. Complexity, leadership, and management in healthcare organisations. *Bmj*, 323(7315):746–749, 2001.

[98] P.E. Plsek. Collaborating across organizational boundaries to improve the quality of care. *American journal of infection control*, 25(2):85–95, 1997.

[99] B.C. Poulton and M.A. West. Effective multidisciplinary teamwork in primary health care. *Journal of Advanced Nursing*, 18(6):918–925, 2008.

[100] M.L. Raymer, W.F. Punch, E.D. Goodman, L.A. Kuhn, and A.K. Jain. Dimensionality reduction using genetic algorithms. *Evolutionary Computation, IEEE Transactions on*, 4(2):164–171, 2000.

[101] Colin Reeves and Christine Wright. An experimental design perspective on genetic algorithms. In *Foundations of Genetic Algorithms 3*, 1995.

[102] Colin R Reeves. Experiments with tuneable fitness landscapes. In *Parallel Problem Solving from Nature PPSN VI*, pages 139–148. Springer, 2000.

[103] Colin R Reeves and Christine C Wright. Epistasis in genetic algorithms: An experimental design perspective. In *Proceedings of the 6th International Conference on Genetic Algorithms*, pages 217–224. Morgan Kaufmann Publishers Inc., 1995.

[104] Ian Reid. Complexity science: Let them eat complexity: the emperor's new toolkit. *BMJ: British Medical Journal*, 324(7330):171, 2002.

[105] D.N. Reshef, Y.A. Reshef, H.K. Finucane, S.R. Grossman, G. McVean, P.J. Turnbaugh, E.S. Lander, M. Mitzenmacher, and P.C. Sabeti. Detecting novel associations in large data sets. *science*, 334(6062):1518–1524, 2011.

[106] A.D. Rodrigues. *Drug-drug interactions*. Informa Healthcare, New York, NY, 2008.

[107] J.A. Rogowski, J.D. Horbar, P.E. Plsek, L.S. Baker, J. Deterding, W.H. Edwards, J. Hocker, A.D. Kantak, P. Lewallen, W. Lewis, et al. Economic implications of neonatal intensive care unit collaborative quality improvement. *Pediatrics*, 107(1):23, 2001.

[108] J.A. Rogowski, J.D. Horbar, D.O. Staiger, M. Kenny, J. Carpenter, and J. Geppert. Indirect vs direct hospital quality indicators for very low-birth-weight infants. *JAMA: the journal of the American Medical Association*, 291(2):202, 2004.

[109] J.A. Rogowski, D.O. Staiger, and J.D. Horbar. Variations in the quality of care for very-low-birthweight infants: implications for policy. *Health Affairs*, 23(5):88–97, 2004.

[110] Jani Rönkkönen, Xiaodong Li, Ville Kyrki, and Jouni Lampinen. A framework for generating tunable test functions for multimodal optimization. *Soft Computing*, 15(9):1689–1706, 2011.

[111] Peter M Rothwell. External validity of randomised controlled trials:to whom do the results of this trial apply?. *The Lancet*, 365(9453):82–93, 2005.

[112] William Rowe, Mark Platt, David C Wedge, Philip J Day, Douglas B Kell, and Joshua Knowles. Analysis of a complete dna–protein affinity landscape. *Journal of The Royal Society Interface*, 7(44):397–408, 2010.

[113] Bill Runciman and Merrilyn Walton. *Safety and ethics in healthcare: a guide to getting it right*. Ashgate Publishing, Ltd., 2007.

[114] E. Salas, N.J. Cooke, and M.A. Rosen. On teams, teamwork, and team performance: Discoveries and developments. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 50(3):540–547, 2008.

[115] T. Sandmann and M. Boutros. Screens, maps & networks: From genome sequences to personalized medicine. *Current Opinion in Genetics & Development*, 22:36–44, 2012.

[116] Elad Schneidman, Susanne Still, Michael J Berry, William Bialek, et al. Network information and connected correlations. *Physical review letters*, 91(23):238701, 2003.

[117] Matthias Schonlau, Rita Mangione-Smith, Kitty S Chan, Joan Keesey, Mayde Rosen, Thomas A Louis, Shin-Yi Wu, and Emmett Keeler. Evaluation of a quality improvement collaborative in asthma care: does it improve processes and outcomes of care? *The Annals of Family Medicine*, 3(3):200–208, 2005.

[118] L.M.T. Schouten, R.P.T.M. Grol, and M.E.J.L. Hulscher. Factors influencing success in quality-improvement collaboratives: Development and psychometric testing of an instrument. *Implementation Science*, 5(1):1–9, 2010.

[119] L.M.T. Schouten, M.E.J.L. Hulscher, J.J.E. Everdingen, R. Huijsman, and R.P.T.M. Grol. Evidence for the impact of quality improvement collaboratives: Systematic review. *Bmj*, 336(7659):1491–1494, 2008.

[120] Loes MT Schouten, MEJL Hulscher, Jannes JE Van Everdingen, Robbert Huijsman, Louis W Niessen, and RPTM Grol. Short-and long-term effects of a quality improvement collaborative on diabetes management. *Implement Sci*, 5:94, 2010.

[121] Paul G Shekelle, Peter J Pronovost, Robert M Wachter, Stephanie L Taylor, Sydney M Dy, Robbie Foy, Susanne Hempel, Kathryn M McDonald, John Ovretveit, Lisa V Rubenstein, et al. Advancing the science of patient safety. *Annals of Internal Medicine*, 154(10):693–696, 2011.

[122] Stephen M Shortell, Jill A Marsteller, Michael Lin, Marjorie L Pearson, Shin-Yi Wu, Peter Mendel, Shan Cretin, and Mayde Rosen. The role of perceived team effectiveness in improving chronic illness care. *Medical care*, 42(11):1040–1048, 2004.

[123] Leif I Solberg. If youve seen one quality improvement collaborative. *The Annals of Family Medicine*, 3(3):198–199, 2005.

[124] Kenneth Tan, Gordon Baxter, Simon Newell, Steve Smye, Peter Dear, Keith Brownlee, and Jonathan Darling. Knowledge elicitation for validation of a neonatal ventilation expert system utilising modified delphi and focus group techniques. *International journal of human-computer studies*, 68(6):344–354, 2010.

[125] Reiko Tanese. *Distributed genetic algorithms for Function Optimization*. PhD thesis, The University of Michigan, Ann Arbor, MI, 1989.

[126] S.L. Taylor, S. Dy, R. Foy, et al. What context features might be important determinants of the effectiveness of patient safety practice interventions? *BMJ Quality & Safety*, 20(7):611–617, 2011.

[127] Dirk Thierens. The linkage tree genetic algorithm. In *Parallel Problem Solving from Nature, PPSN XI*, pages 264–273. Springer, 2010.

[128] Dirk Thierens and Peter AN Bosman. Hierarchical problem solving with the linkage tree genetic algorithm. In *Proceeding of the fifteenth annual conference on Genetic and evolutionary computation conference*, pages 877–884. ACM, 2013.

[129] Nicholas Tomko, Inman Harvey, and Andrew Philippides. Unconstrain the population: The benefits of horizontal gene transfer in genetic algorithms. In *SmartData*, pages 117–127. Springer, 2013.

[130] Shaun Treweek and Merrick Zwarenstein. Making trials matter: pragmatic and explanatory trials and the problem of applicability. *Trials*, 10(37):9, 2009.

[131] M.E. Turner. *Groups at work: Theory and research*. Lawrence Erlbaum, Hillsdale, NJ, 2000.

[132] Ryan J Urbanowicz and Jason H Moore. The application of michigan-style learning classifiersystems to address genetic heterogeneity and epistasisin association studies. In *Proceedings of the 12th annual conference on Genetic and evolutionary computation*, pages 195–202. ACM, 2010.

[133] G. Vaidyanathan. Redefining clinical trials: The age of personalized medicine. *Cell*, 148(6):1079–1080, 2012.

[134] Vesselin K Vassilev, Terence C Fogarty, and Julian F Miller. Information characteristics and the structure of landscapes. *Evolutionary Computation*, 8(1):31–60, 2000.

[135] Edward Weinberger. Correlated and uncorrelated fitness landscapes and how to tell the difference. *Biological cybernetics*, 63(5):325–336, 1990.

[136] John Wennberg and Alan Gittelsohn. Small area variations in health care delivery: A population-based health information system can guide planning and regulatory decision-making. *Science*, 182(4117):1102–1108, 1973.

[137] John E Wennberg. *Tracking Medicine: A Researcher's Quest to Understand Health Care*. Oxford University Press, USA, 2010.

[138] M.A. West. *Effective teamwork: Practical lessons from organizational research*. Blackwell Publishing, Oxford, 2012.

[139] Tim Wilson, Donald M Berwick, and Paul D Cleary. What do collaborative improvement projects do? experience from seven countries. *Joint Commission Journal on Quality and Patient Safety*, 29(2):85–93, 2003.

[140] Tim Wilson, Tim Holt, and Trisha Greenhalgh. Complexity and clinical care. *Bmj*, 323(7314):685–688, 2001.

[141] David D Woods, Leila J Johannesen, Richard I Cook, and Nadine B Sarter. Behind human error: Cognitive systems, computers and hindsight. Technical report, DTIC Document, 1994.

[142] David D Woods, Emily S Patterson, and Richard I Cook. Behind human error: taming complexity to improve patient safety. *Handbook of Human Factors and Ergonomics in Health Care and Patient Safety. London: Lawrence Erlbaum*, pages 459–76, 2007.

[143] Michael Wooldridge. *An introduction to multiagent systems*. John Wiley & Sons, 2009.

[144] Alden H Wright, Richard K Thompson, and Jian Zhang. The computational complexity of NK fitness functions. *IEEE Transactions on Evolutionary Computation*, 4(4):373–379, 2000.

[145] S. Wright. The roles of mutation, inbreeding, crossbreeding and selection in evolution. In *Proceedings of the sixth international congress on genetics*, volume 1, pages 356–366, 1932.

[146] Paul C Young, Gordon B Glade, Gregory J Stoddard, and Chuck Norlin. Evaluation of a learning collaborative to improve the delivery of preventive services by pediatric practices. *Pediatrics*, 117(5):1469–1476, 2006.

[147] Z.J. Yu, F. Haghighat, B. Fung, and L. Zhou. A novel methodology for knowledge discovery through mining associations between building operational data. *Energy and Buildings*, 2011.

[148] J.A.F. Zupancic, D.K. Richardson, J.D. Horbar, J.H. Carpenter, S.K. Lee, G.J. Escobar, et al. Revalidation of the score for neonatal acute physiology in the vermont oxford network. *Pediatrics*, 119(1):e156–e163, 2007.

Chapter 4

NM Landscapes

4.1 Abstract

For the past 25 years, *NK* landscapes have been a classic benchmark for modeling combinatorial fitness landscapes with epistatic interactions between up to $K + 1$ of N binary features. However, the ruggedness of *NK* landscapes is only tunable in large discrete jumps, and computing the global optimum of unrestricted NK landscapes is an NP-complete problem. Furthermore, the range of fitness values can vary widely between different landscapes, but since this range of fitnesses is unknown one cannot properly normalize fitnesses, as is necessary for fair comparisons of fitness across different random landscapes. Walsh polynomials are a superset of NK landscapes that avoid some of these problems, but both Walsh polynomials and NK landscapes are only defined on binary alphabets and their representation of epistatic interactions is not intuitive. In this paper, we propose a new class of benchmarks called *NM* landscapes, where M refers to the Maximum order of epistatic interactions between N features. Like Walsh polynomials, *NM* landscapes are much more smoothly tunable in ruggedness than *NK* landscapes. For all NM landscapes the location

and the value of the global optimum is trivially known. For a subset of *NM* landscapes the location and the value of the global minimum is also known, enabling proper normalization of fitnesses. NM landscapes use a natural and transparent representation of epistasis and work with alphabets of any arity, from binary to real-valued. We discuss several advantages of *NM* landscapes as benchmark problems for evaluating search strategies.

Simulated landscapes are widely used for evaluating search strategies, where the goal is to find the landscape location with maximum fitness value [62] [27]. Without loss of generality and for notational simplicity, we assume function maximization, rather than minimization, throughout this paper.

NK Landscapes [62] have been classic benchmarks for generating landscapes with epistatic interactions. They are described by two parameters: *N* specifies the number of binary features and K specifies that the maximum degree of epistatic interactions among the features is $K + 1$ [60]. *NK* landscapes have been used in many applications (e.g., [2, 36, 82, 112, 129]) and widely studied in theory (e.g., [15, 53, 69, 92, 144]), as they can generate landscapes with tunable ruggedness by varying K. However, NK landscapes have several limitations. Buzas and Dinitz [15] recently showed that the expected number of local peaks in *NK* landscapes rises in large discrete jumps as K is increased, but actually decreases as a function of the number of interaction terms for a given K (Fig. 1, red lines). Additionally, the problem of finding the location and value of the global optimum of unrestricted *NK* landscapes with $K > 1$ is NP-complete [144] (although for restricted classes one can use dynamic programming [144] [35] or approximation algorithms [144]). *NK* landscapes have only been defined for binary alphabets.

Walsh polynomials are a superset of NK landscapes that overcome some of the limitations of NK landscapes. For example, they allow more explicit control over which interaction terms are present. The problem of finding the global maximum value of a Walsh polynomial is also NP-complete, although a restricted subset of Walsh polynomials has a known global maximum [125]. However, even in this case finding the global minimum is still NP-complete, preventing proper normalization by the range of fitnesses in the landscapes. As with NK landscapes, Walsh polynomials are only defined for binary alphabets.

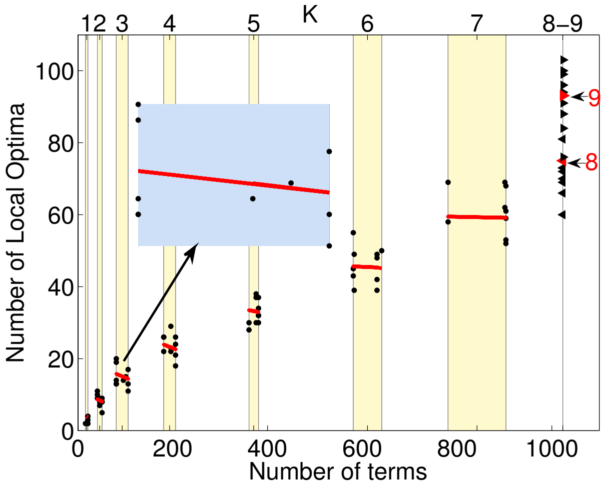


Figure 4.1: Number of local peaks in NK landscapes with $N = 10$, as a function of the number of terms in the equivalent parametric interaction model (m, bottom x-axis) for $K = \{1, 2, ..., 9\}$ (top x-axis). The black dots show empirical results of 10 random landscapes generated for each value of K; red lines show the expected number of peaks (L) of these same landscapes computed according to the formula given in [15]. The inset shows a magnification of the $K = 3$ results.

In this paper, we introduce a different, more flexible subset of general interaction models that we dub *NM* landscapes. Like *NK* landscapes and Walsh polynomials, *NM* landscapes incorporate epistatic feature interactions. However, NM landscapes also (a) include epistasis in a natural and transparent manner, (b) have known value and location of the maximum fitness, (c) work with alphabets of any arity, including discrete and real-valued alphabets, (d) with additional constraints have known value and location of the minimum fitness, and (e) when coefficients are chosen properly, have relatively smoothly tunable ruggedness. In Section 4.2 we introduce the general class of parametric interaction models and Walsh polynomials, then in Section 4.3 we define NM landscapes and prove the properties (a),(b),(c) and (d) above. In Sections 4.4 and 4.5 we describe experiments and results that demonstrate property (e) above. In Section 4.6 we discuss the importance of these properties and point out several advantages of NM landscapes as benchmark problems for studying search in tunably rugged landscapes.

4.2 Interaction Models and Walsh Polynomials

Walsh polynomials provide a mathematical framework for defining any real-valued function on bit strings [31] [59]. A Walsh polynomial has the following form:

$$f(\mathbf{y}) = \sum_{j=0}^{2^q-1} \omega_j \psi_j(\mathbf{y}) \tag{4.1}$$

where q is the length of the bit string $\mathbf{y}$, each bit $y_i \in \{0, 1\}$, and each $\omega_j \in \mathbb{R}$. The Walsh function $\psi_j(\mathbf{y})$ corresponding to the jth partition is defined as:

$$\psi_j(\mathbf{y}) = \begin{cases} 1, & \text{if } \mathbf{y} \wedge j_2 \text{ has even parity} \\ -1, & \text{otherwise} \end{cases} \tag{4.2}$$

where j_2 denotes the binary representation of j.

NK landscapes are a subset of Walsh polynomials. Walsh polynomials have a one-to-one correspondence with the more general class of general parametric interaction models, when such models are restricted to binary alphabets [59].

A fitness landscape F can be defined for N features using a general parametric interaction model of the form:

$$F(\mathbf{x}) = \sum_{k=1}^{m} \beta_{U_k} \prod_{i \in U_k} x_i \tag{4.3}$$

where m is the number of terms, and each of m coefficients $\beta_{U_k} \in \mathbb{R}$. For $k = 1 \ldots m$, $U_k \subseteq \{1, 2, \ldots, N\}$, where U_k is a set of indices of the features in the kth term, and the length $|U_k|$ is the order of the interaction. We adopt the convention that when $U_k = \emptyset$, $\prod_{j \in U_k} x_j \equiv 1$, so β_0 represents the mean value of the landscape. If the parametric interaction model is defined on a binary alphabet, we adopt the convention that binary values are represented as $x_i \in \{-1, 1\}$ (rather than $\{0, 1\}$, as in Walsh polynomials). However, general parametric interaction models are also well defined for discrete valued features with higher arities as well as for real-valued alphabets and provide a more intuitive way of representing epistatic interactions among features.

A more readable notation for Eq. (4.3) is as follows:

$$\begin{aligned} F(\mathbf{x}) = \beta_0 + &\sum_{i=1}^{N} \beta_i x_i + \\ + &\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \beta_{i,j} x_i x_j + \\ + &\sum_{i=1}^{N-2} \sum_{j=i+1}^{N-1} \sum_{k=j+1}^{N} \beta_{i,j,k} x_i x_j x_k + H \end{aligned} \tag{4.4}$$

where we only explicitly show up to third order interactions and H represents higher order interactions up to some maximum order $M \leq N$. Note that some β_{U_k} parameters may be zero, so not all terms need be present.

For example, consider a model with $N = 2$ loci and $U_1 = \emptyset$, $U_2 = \{1\}$, $U_3 = \{2\}$ and $U_4 = \{1, 2\}$. The interaction model for this example is:

$$F(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{1,2} x_1 x_2 \tag{4.5}$$

where β_0 is the average value of all fitnesses in the landscape, β_1 and β_2 are the coefficients of the main effects of the binary features x_1 and x_2, and $\beta_{1,2}$ is the coefficient of the second order epistatic interaction $x_1 x_2$. The Walsh polynomial corresponding to Eq. (4.5) is:

$$\begin{aligned} f(\mathbf{y}) &= \omega_0 \psi_0(\mathbf{y}) + \omega_1 \psi_1(\mathbf{y}) + \omega_2 \psi_2(\mathbf{y}) + \omega_3 \psi_3(\mathbf{y}) \\ &= \beta_0 \psi_0(\mathbf{y}) - \beta_1 \psi_1(\mathbf{y}) - \beta_2 \psi_2(\mathbf{y}) + \beta_{1,3} \psi_3(\mathbf{y}) \end{aligned} \tag{4.6}$$

where

$$y_i = \begin{cases} 1, & \text{when } x_i = 1 \\ 0, & \text{when } x_i = -1 \end{cases} \tag{4.7}$$

Notice that there is a one-to-one correspondence of each term in Eq. (4.5) with each term in Eq. (4.6) but the signs of the coefficients are different. Specifically, for the example above:

$$\beta_0 = \omega_0, \;\; \beta_1 = -\omega_1, \;\; \beta_2 = -\omega_2, \;\; \beta_{1,2} = \omega_3 \tag{4.8}$$

A random point selected in the search space of a Walsh polynomial can be forced to be the global maximum by properly adjusting the sign of each of the non-zero Walsh coefficients, with the maximum fitness value equal to the sum of the absolute values of all Walsh coefficients [125]. However, the location and the value of the global minimum is still unknown.

General parametric interaction models are the standard models used in statistics to study effects of multiple features on an outcome (e.g., [80]). They are easy to define and the interactions are transparent and easy to interpret (unlike in *NK* landscapes and Walsh polynomials). For example, the interaction terms present in Eq. (4.5) are clearly evident, whereas the Walsh functions ψ_i in Eq. (4.6) obscure this. To date, general parametric interaction models have received very little attention in the evolutionary computation literature, with notable exceptions [101–103].

In [15] the authors show that for every *NK* landscape with a given K, one can create an equivalent parametric interaction model, where the maximum order of interactions is $K+1$. They show that the *NK* algorithm dictates that the interaction model contain all main effects and sub-interactions contained in higher order interactions. For example, if a non-zero interaction coefficient $\beta_{i,j,k}$ is present in an NK landscape, then there will generally be non-zero coefficients $\beta_i, \beta_j, \beta_k, \beta_{i,j}, \beta_{i,k}, \beta_{j,k}$ (there is an infinitesimally small probability that one or more of these coefficients may be zero). For the classic *NK* model where K is constant and $K \ll N$, main effect coefficients have the largest expected magnitude, second order interactions have larger expected magnitude than third order interactions, and so on [15]. Thus, *NK* landscapes are a very restricted subset of Walsh polynomials and the more general class of parametric interaction models.

4.3 NM Landscapes

The class of Walsh polynomials is a subset of the larger class of general interaction models. Here we introduce a different subset of general interaction models called *NM* landscapes, where N is the number of features and all interactions in the model are of order $\leq M$.

Definition 1: NM models comprise the set of all general interaction models specified by Eq. (4.3), with the added constraints that (a) all coefficients β_{U_k} are non-negative, (b) each feature value x_i ranges from negative to positive values, and (c) the absolute value of the lower bound of the range $\leq$ the upper bound of the range of x_i.

In this work, each β_{U_k} is randomly created as follows:

$$\beta_{U_k} = e^{-abs(\mathbb{N}(0,\sigma))} \tag{4.9}$$

where $\mathbb{N}(0, \sigma)$ is a random number drawn from a Gaussian distribution with 0 mean and standard deviation of σ, which results in fitnesses that are symmetric around 0 (Fig. 4.2). As the value of σ increases, the means and standard deviations of the coefficients decrease, which results in a smaller range of fitness values and increasing clumping of fitness values (Fig. 4.2). In contrast, when coefficients are drawn from a uniform distribution in the range $[0, 1]$, the fitnesses are skewed left (Fig. 4.3). NM landscapes offer several desirable properties, as described in the following.

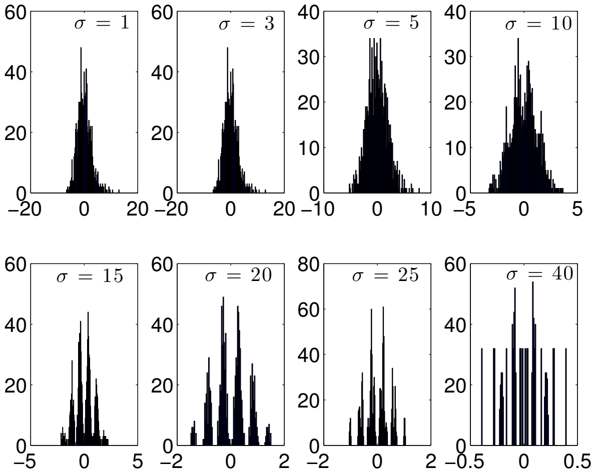


Figure 4.2: Histograms of all 1024 fitnesses in NM landscapes for $M = 2$, $N = 10$ and coefficients drawn from Eq. (4.9) with σ as indicated.

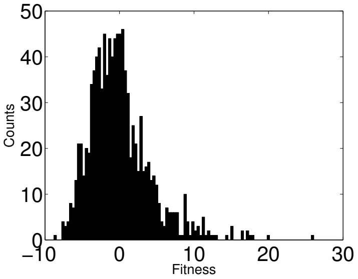


Figure 4.3: Histograms of all 1024 fitnesses in binary NM landscapes for $M = 2$, $N = 10$ and coefficients drawn from a uniform distribution in the range $[0, 1]$.

Proposition 1: NM landscapes with a binary alphabet have a known global maximum.

Proof. By Definition 1, $\beta_{U_k} > 0$ for all non-zero terms. Thus, the maximum possible value for each term ($\beta_{U_k} \prod_{j \in U_k} x_j$) in an *NM* landscape with a binary alphabet $x_i \in \{-1, 1\}$ is achieved when:

$$x_i = 1, \quad \forall i = 1 \dots n \tag{4.10}$$

and the value of the global maximum is:

$$
\begin{aligned}
F_{max} = \beta_0 + \sum_{i=1}^{N} \beta_i + \\
+ \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \beta_{i,j} + \\
+ \sum_{i=1}^{N-2} \sum_{j=i+1}^{N-1} \sum_{k=j+1}^{N} \beta_{i,j,k} + \sum_{\forall \beta_{U_h}} \beta_{U_h}
\end{aligned}
\tag{4.11}
$$

where β_{U_h} are the coefficients of all the remaining higher order interactions. Note that the calculation of Eq. (4.11) has time complexity of $O(m)$, where m is the number of terms in the model. ■

Proposition 2: *NM* landscapes can be defined on discrete alphabets of any arity or on real-valued alphabets, and the value and location of a global maximum is independent of the discretization of the alphabet.

Proof. By Definition 1, all coefficients are non-negative, therefore the maximum F_{max} of an *NM* landscape with any discrete or real-valued alphabet $x_{a,b}$ defined in the range $[-a, b]$ where $a \leq b$, occurs when $x_i = b, \forall i = 1 \ldots N$ and its value is:

$$
\begin{aligned}
F_{max} = \beta_0 + \sum_{i=1}^{N} \beta_i b + \\
+ \sum_{i=1}^{(N-1)} \sum_{j=i+1}^{N} \beta_{i,j} b^2 + \\
+ \sum_{i=1}^{(N-2)} \sum_{j=i+1}^{(N-1)} \sum_{k=k+1}^{N} \beta_{i,j} b^3 + \sum_{\forall \beta_{U_h}} \beta_{U_h} b^{|U_h|}
\end{aligned}
\tag{4.12}
$$

where β_{U_h} are the coefficients of all the remaining higher order interactions and the lengths $|U_h|$ are the orders of these interactions. Thus, the magnitude of F_{max} is a function of b, but is independent of the arity of the alphabet. ■

We note that it is trivial to extend this proposition and proof to NM landscapes with heterogeneous alphabets (i.e., different ranges and/or arities for each feature variable), as long as the lower bound for each feature is negative, the upper bound is positive, and the absolute value of the lower bound is $\leq$ the upper bound. However, for notational simplicity we only demonstrate the proof for homogeneous alphabets. We refer to the above described general NM landscapes as Type I NM landscapes.

We conjecture that changing the arity of features in NM landscapes does not change the number, locations, or values of the local peaks or global minima, because higher arity alphabets simply sample the same landscape at a higher resolution that interpolates between the local peaks. (Empirical data, not shown, supports this conjecture.)

Note that interaction models with all non-negative interaction coefficients β_{U_k}, but no negative feature values, generate unimodal landscapes. However, since alphabets in NM landscapes are defined to include both negative and positive features, NM landscapes have multiple local optima whenever any interaction terms are included.

Proposition 3: NM landscapes that include all main effects and any odd order interactions have exactly one global maximum.

Proof. By proposition 1, a maximum fitness F_{max} of an *NM* landscape is achieved at point $\mathbf{x} = [b, b, \ldots b]$. Let's assume there exists another point $\mathbf{y} = [y_1, y_2, \ldots y_n]$, where $\mathbf{y} \neq \mathbf{x}$ (i.e, there exists at least one i such that $y_i \neq b$), that is also a global maximum. Two cases must be considered. In the first case, if we assume that $\exists i$, $abs(y_i) < b$, since each $x_i \in [-a, b]$ and $a \leq b$ (by Definition 1) then the value of the interaction model at point

y will be strictly less than the value at point **x**. In the second case, if we assume that $\exists i, y_i = -b$, all even order terms will have the same values at points **x** and **y** but any odd order terms (including main effects) containing y_i will be negative (since, by Definition 1, all coefficients are non-negative), therefore the overall sum of terms in **y** will be strictly less than the global maximum achieved at point **x**. Thus, by contradiction, **x** is the only global maximum. ■

Proposition 4: NM Landscapes that include only even order terms with alphabets in the range [-b, b] are symmetric and have exactly two global maxima at maximal distance apart in feature space.

Proof. Since $x_i^{2t} = (-x_i)^{2t}$, $\forall t \in \mathbb{I}$, then for all NM landscapes with only even order terms, $F(\mathbf{x}) = F(-\mathbf{x})$ for each pair of points $\mathbf{x} = [x_1, x_2, \ldots, x_N]$ and $-\mathbf{x} = [-x_1, -x_2, \ldots, -x_N]$. Thus, NM landscapes with only even order interactions and alphabets in $[-b, b]$ range are symmetric and the two global maxima are at locations $[b, b, \ldots b]$ and $[-b, -b, \cdots - b]$, which are the maximum distance away from each other in the feature space. ■

When the value of the global maximum of the landscape is known one can partially normalize fitnesses to the range ≤ 1 using the following formula:

$$F = \frac{F}{F_{max}} \tag{4.13}$$

However, proper normalization of fitnesses to the interval [0,1] also requires prior knowledge of the global minimum of the landscape, as follows:

$$F = \frac{F - F_{min}}{F_{max} - F_{min}} \tag{4.14}$$

To this end, we define subsets of NM landscapes that have a known global minimum. While there are many ways to do this, below we present two such subsets.

Proposition 5: NM landscapes that include only main effects with odd indices (e.g., x_1, x_3, x_5, etc.) and any terms with an odd number of odd indices (e.g., x_1x_2, $x_1x_3x_5$, $x_1x_3x_6x_7$, etc.) and alphabets in the range $[-1, 1]$ have a global minimum located at point $[-1, 1, -1, 1 \ldots]$. For example, models of this form including up to $M = 3$ order terms are given by:

$$\begin{aligned} F(x) =& \beta_0 + \sum_{i\ odd}^{N} \beta_i x_i + \sum_{\substack{i\ odd,\\ j\ even}}^{N} \beta_{i,j} x_i x_j + \\ & \sum_{\substack{i\ odd,\\ j\ even,\\ k\ even}}^{N} \beta_{i,j,k} x_i x_j x_k + \sum_{\substack{i\ odd,\\ j\ odd,\\ k\ odd}}^{N} \beta_{i,j,k} x_i x_j x_k \end{aligned} \tag{4.15}$$

Proof. At the point $[-1, 1, -1, 1 \ldots]$ all terms with an odd number of odd indices will have a negative sign, as the product of an odd number of negative numbers is negative. Thus, this point is the global minimum of the landscape with value:

$$F_{min} = -\big(\beta_0 + \sum_{i\ odd}^{N} \beta_i + \sum_{\substack{i\ odd,\\ j\ even}}^{N} \beta_{i,j} + \ldots\big) \tag{4.16}$$

(where only terms up through second order are explicitly shown above). ■

We refer to the NM landscapes defined in Proposition 5 as Type II NM landscapes. Note that Type II NM landscapes can easily be extended to alphabets in the range $[-a, b]$, where $a \leq b$, although for notational simplicity we have limited the range to $[-1, 1]$ in the above.

Proposition 6: NM landscapes with only odd order terms and alphabets in the range $[-a, b]$, where $a \leq b$, have a global minimum located at $[-a, -a, \ldots, -a]$.

Proof. By Definition 1, $\beta_{U_k} > 0$ for all non-zero terms, $x_i \in [-a, b]\ \forall i$, and $a \leq b$. Therefore the value of each term $T_k = \beta_{U_k} * x_{U_k}$ has to be $\geq -|a|^{|U_k|}$. When all the features $x_i = -a$, $T_k = -|a|^{|U_k|}$. Therefore the point $[-a, -a, \ldots, -a]$ is a global minimum with value:

$$\begin{aligned} F_{min} = -(\beta_0 + \sum_{i=1}^{N} \beta_i a + \\ + \sum_{i=1}^{(N-2)} \sum_{j=i+1}^{(N-1)} \sum_{k=k+1}^{N} \beta_{i,j} a^3 + \sum_{\forall \beta_{U_h}} \beta_{U_h} a^{|U_h|}) \end{aligned} \tag{4.17}$$

■

We refer to the NM landscapes defined in Proposition 6 as Type III NM landscapes. When $a = b$ the global maximum and minimum of Type III NM landscapes have the same absolute value, but opposite signs. Because NM landscapes allow only non-negative coefficients but require both positive and negative feature values, we are thus able to construct NM landscapes with known maximum and known minimum, enabling normalization of fitnesses to the range [0,1] by equation (4.14). In contrast, Walsh polynomials allow both positive and negative coefficients, but have only non-negative feature values. Thus, while it is possible to manipulate the signs of the Walsh coefficients to specify the location of the global maximum [125], the global minimum of Walsh polynomial is still unknown, even if one restricts the order of the interactions as in Type II or Type III NM landscapes.

4.4 Experiments

4.4.1 Ruggedness

We illustrate how ruggedness changes on binary *NM* landscapes with coefficients drawn from the distribution in Eq. (4.9). Since we assess the ruggedness of these models using exhaustive search, we limit our experiments to $N \leq 15$.

In one set of experiments, we generated a random Type I "master" *NM* model, including terms for all N main effects and the $\sum_{i=1}^{N} \binom{N}{i}$ possible interaction terms (e.g., for $N = 10$ there are 1023 overall terms; 1013 interaction terms plus 10 main effects). We then systematically created subsets of the master model that include an increasing number m of terms from the master model, as follows. We started with a base model that includes all main effects. Random second order terms were then added in groups of 10 (or less if there are not 10 left). After we had included all of the second order terms, we began adding randomly selected groups of 10 third order terms, and so on, until the single N-order interaction term was included. We performed repetitions of these incremental explorations of 100 master NM models for each of $N = 10$ with $\sigma = 10$, and $N = 15$ with $\sigma \in \{15, 20, 100\}$.

In another set of experiments, we similarly created 100 master Type II NM landscapes according to Eq. (4.15), with $N = 10$ and $\sigma = 10$. We created increasing subsets from the master models, as described above for the Type I landscapes.

We computed two standard measures of landscape ruggedness [134, 135]: (a) we counted the number of local peaks (where a local peak is defined as any point whose fitness value is greater than that of all of its neighbors); (b) we computed the lag 1 autocorrelation of random walks through the landscapes.

4.4.2 Distribution of fitnesses and local peaks

We generated representative NM landscapes with $N = 10$ and $\sigma = 10$ for each of $M \in \{1, 2, 3, 4, 6, 10\}$ for both Type I and Type II NM landscapes. We visualize these landscapes by plotting the fitnesses of all points in the landscape as a function of their distances (in feature space) to the global optimum, indicating which are local optima.

4.4.3 Basin of attraction of global optimum

We assessed the size of the basin of attraction of the global maximum of Type I and Type III NM landscapes and NK landscapes for different values of $K = M - 1 \in \{1, 9\}$ and $N = 10$. The fitness matrix of NK landscapes is generated from random uniform numbers in the $[0, 1]$ range. We calculated the size of the global basin of attraction as a weighted sum of the points in the landscape that can reach the global maximum using only hill climbing. Each point was weighted based on the percentage of its immediate neighbors with higher fitnesses that were also in the basin of attraction of the global maximum.

4.4.4 Searchability of the landscapes

We assessed how searchable NM landscapes are using simple genetic algorithms (GAs). In all the experiments we used a GA with $N = 32$, $\sigma = 32$, population size 256, crossover rate 0.7, uniform mutation and the number of random seeds of 32 (these parameter values were selected to be the same as in [125]).

We studied search on Type I NM landscapes with $M = 2$ and $P \in \{0, 0.1, \ldots, 1\}$ proportions of all possible second-order interactions. We studied the search on Type III NM landscapes with $M \in \{1, 3, 5\}$ including all possible main effects and odd order interactions of order $\leq M$.

4.5 Results

The number of local peaks L in NM landscapes increases relatively smoothly as we increase the number of terms (m) in both Type I and Type II NM landscapes (i.e., the regions between the vertical lines on Fig. 4.4) and as we increase the maximum order of interactions M (i.e., as we cross a vertical line on Fig. 4.4).

Note that the average number of local peaks for a given m in both Type I and Type III landscapes is on the same order of magnitude as the expected number of local peaks in NK models with the same N and $K + 1 = M$ (compare Figs. 4.4a and 4.4c to Fig. 1).

Similarly, the lag 1 autocorrelation of random walks through both Type I and Type III NM landscapes decreases relatively smoothly as the number of terms m is increased in models with a given M, as well as when the maximum order of interactions M is increased (Figs. 4.4b and 4.4d), where lower autocorrelation corresponds to greater ruggedness. Notice that, especially for small M, the increase in ruggedness (as measured by both the number of local peaks and the lag 1 autocorrelation) asymptotically slows as the number of terms m increases (Fig. 4.4).

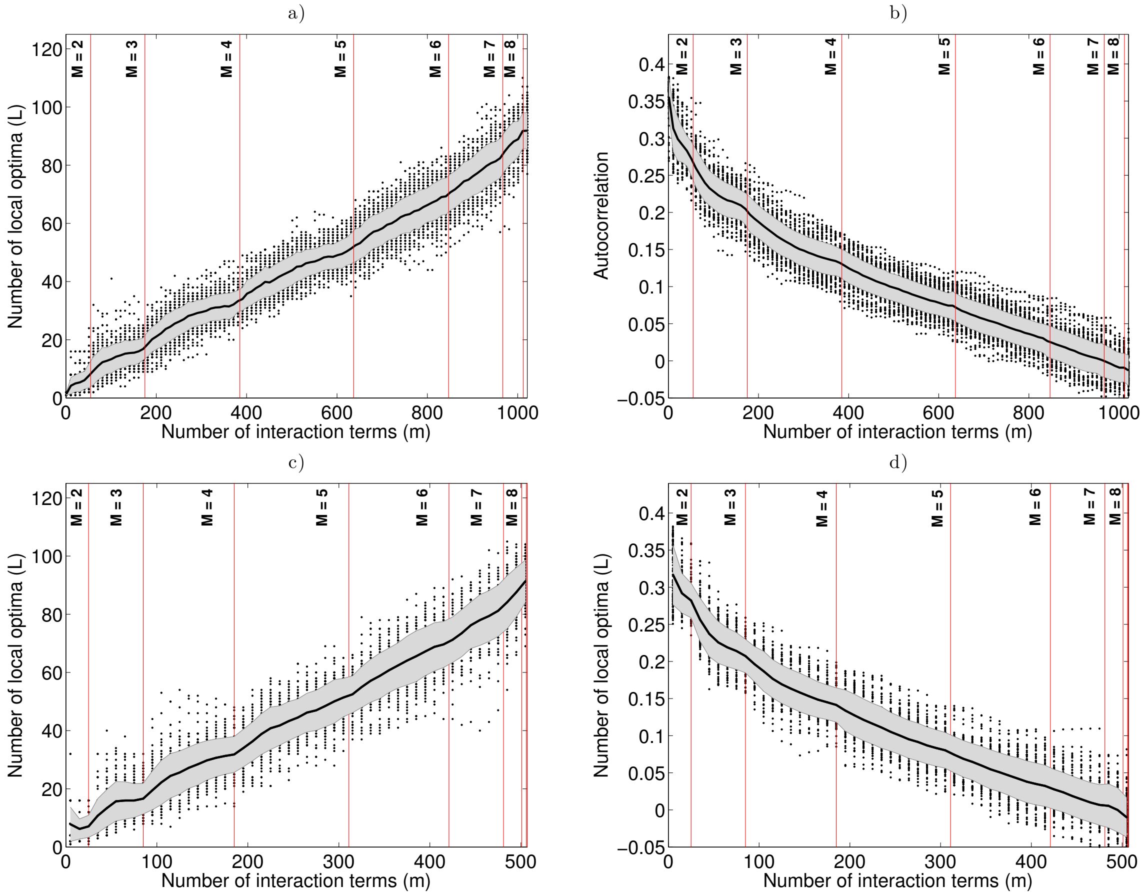


Figure 4.4: Number of local peaks in landscape as we increase the number of terms (x-axis) and order of interactions (labels near top), for NM landscapes with $N = 10$ and $\sigma = 10$. The gray area shows the standard deviation and black line shows the mean for 100 random NM landscapes. a) and c) show the number of local peaks, b) and d) show the lag 1 autocorrelation for the general model and the model with known global minimum respectively.

The larger the σ values in Eq. (4.9), the smaller the range of fitness values in the landscape (Fig. 2), resulting in larger standard deviations of both the number of local peaks in the landscape (shown in Fig. 4.5 for $N = 15$) and the autocorrelation (not shown).

We show the fitnesses of all points in representative NM landscapes with $N = 10$ and $\sigma = 10$ as a function of their distances in feature space to the global maximum for both the

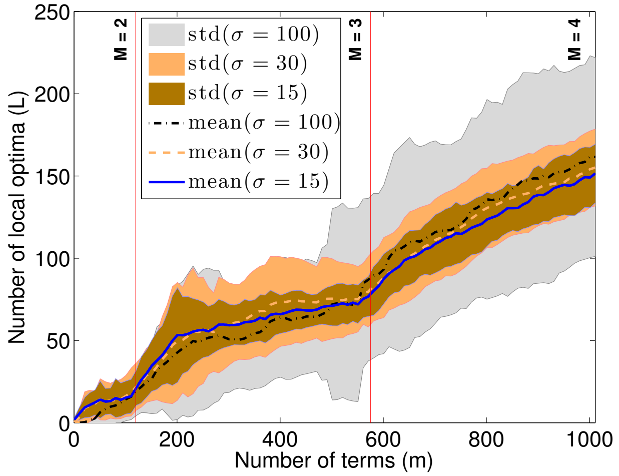


Figure 4.5: Number of local peaks in landscape as we increase the number of terms (x-axis) and order of interactions (labels near top), for NM landscapes with $N = 10$ and $\sigma = 10$. The gray area shows the standard deviation and black line shows the mean for 100 random NM landscapes.

Type I (Fig.4.6) and Type II (Fig. 4.7) NM landscapes. The global maximum is indicated by the leftmost red × in each panel and the remaining red ×'s are sub-optimal local peaks. As we increase the maximum order of interactions M, the fitness difference between the global maximum and other points in the landscape increases; this effect is amplified for Type II NM landscapes (Fig. 4.7) relative to Type I NM landscapes (Fig. 4.6).

In both Type I and Type II NM landscapes the distance in feature space between the global maximum and the nearest local peak generally decreases with increasing M and the sizes of the basin of attraction for the global maximum decreases (Fig. 4.8). Our results show that NK and NM landscapes have similar sizes of the basin of attraction for the global maximum for small and large K = $(M - 1)$. However, the size of the basin of attraction for the global maximum of both Type I and Type II NM landscapes decreases with increasing M rapidly for $M \leq 5$ then levels out, while for NK landscapes the decrease is more gradual (see Fig. 4.8).

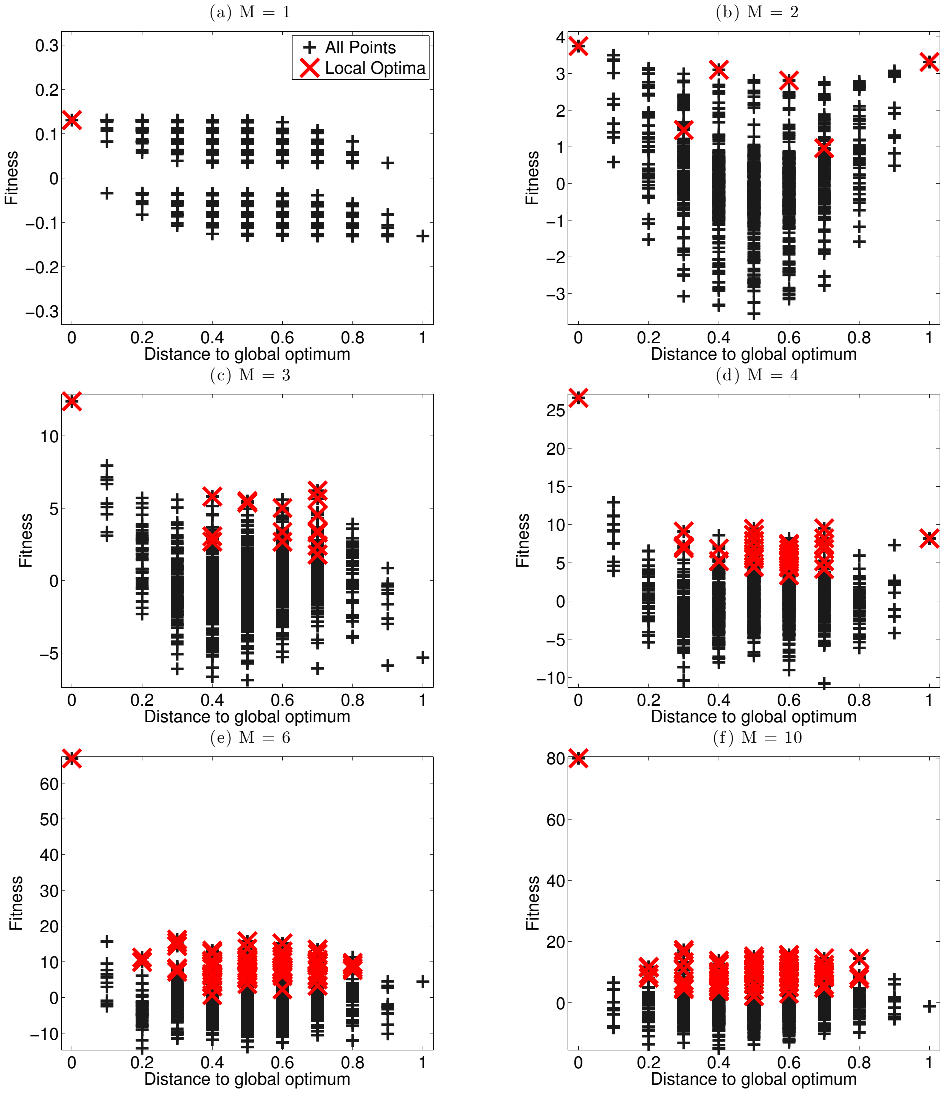


Figure 4.6: Visualization of fitnesses of all the points in representative Type I binary NM landscapes with $N = 10$, $\sigma = 10$ versus their distances from the global optimum in feature space for (a) $M = 1$, (b) $M = 2$, (c) $M = 3$, (d) $M = 4$, (e) $M = 6$, (f) $M = 10$. In these models, all possible interactions for orders up to M were included.

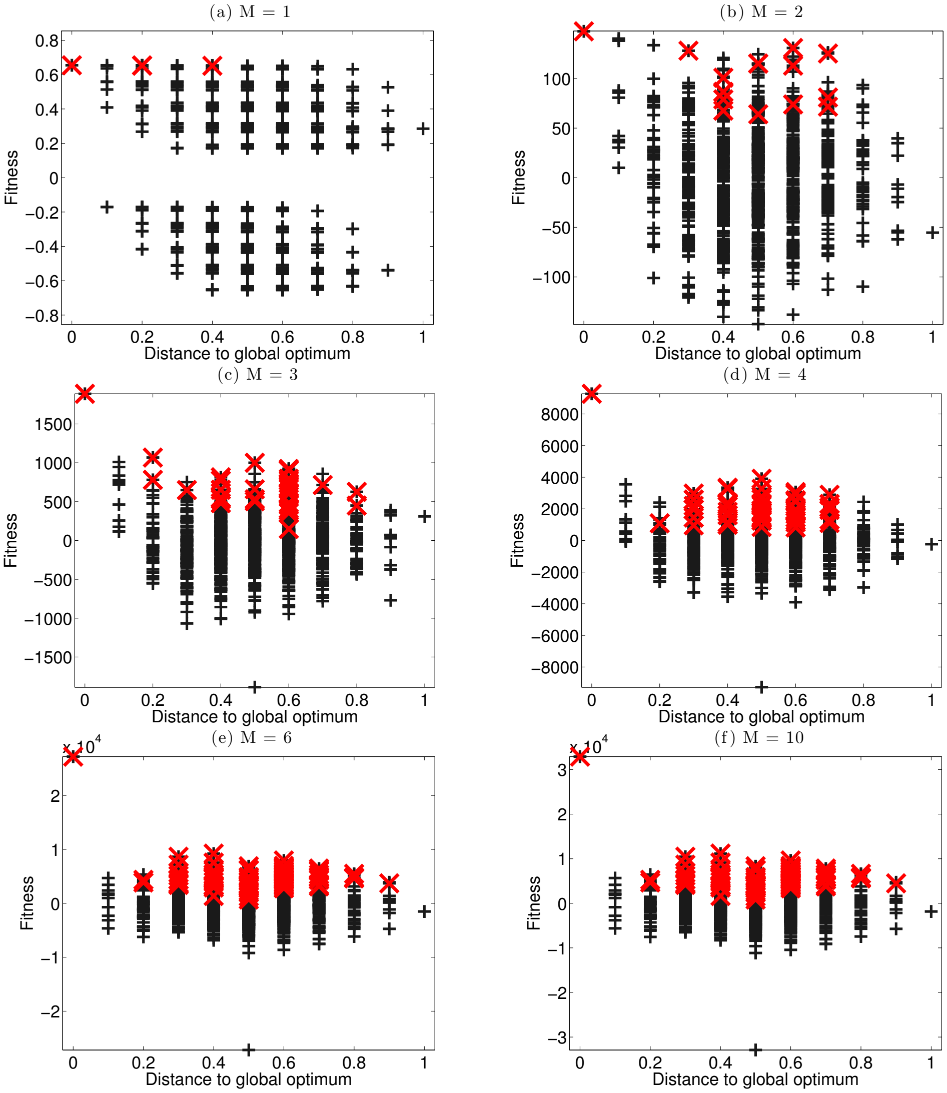


Figure 4.7: Same as Fig. 4.6 for the Type II NM landscapes.

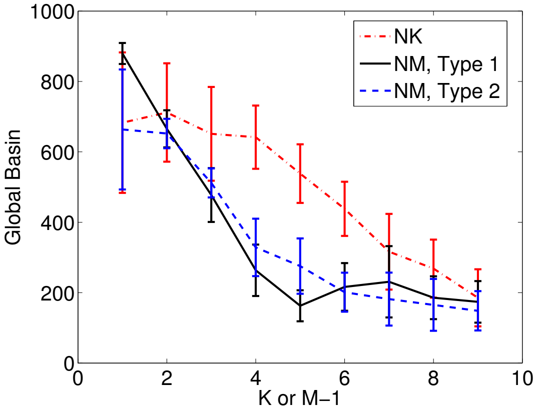


Figure 4.8: The size of the global basin for NK, NM, Type 1 and NM, Type 2 landscapes. Note that $K = M - 1$.

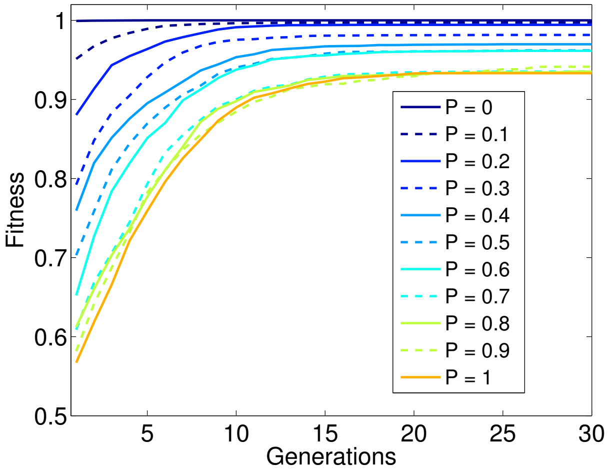


Figure 4.9: Best fitnesses in population of 256 agents shown for 30 generations of search with Genetic Algorithm. Search is performed on 32 random NM landscapes with $M = 2$, $N = 32$ and different proportions of second-order interactions shown in legend. Crossover rate is 0.7.

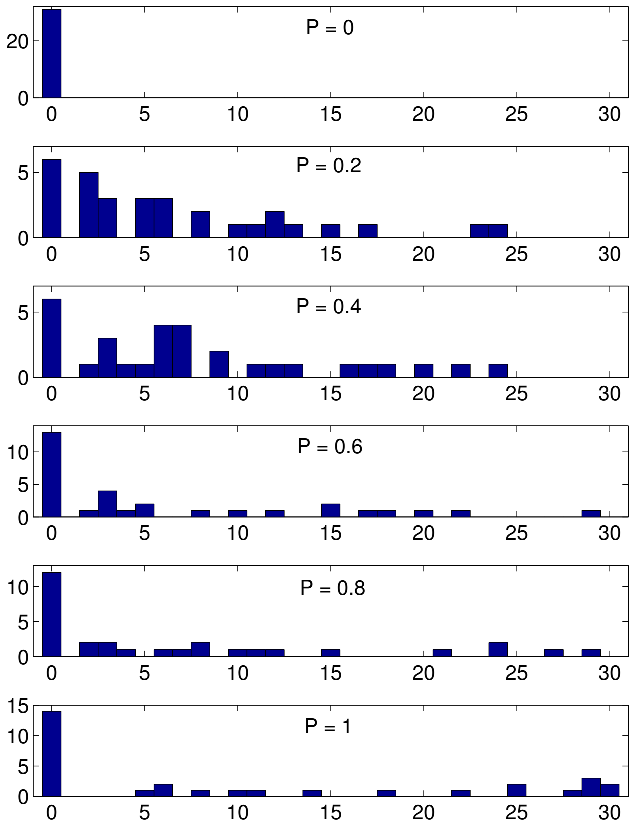


Figure 4.10: Histograms of the distances between the best solution at generation 30 and the global maximum for 32 random NM landscapes with $M = 2$ and $N = 32$. Results are shown for NM landscapes with different proportions of second-order interactions (see histogram titles).

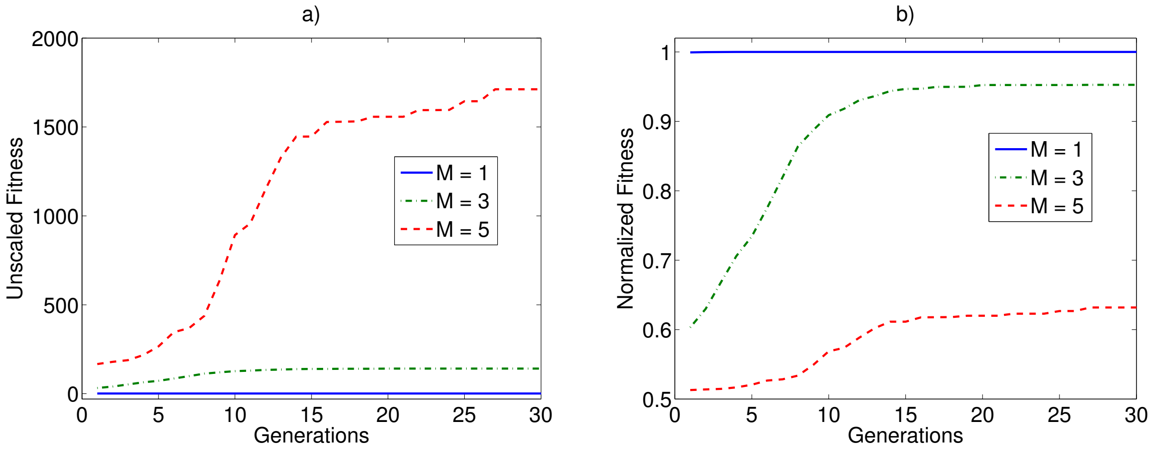


Figure 4.11: Mean of the best fitnesses found by GA over 30 generations (x-axis) for 32 random NM landscapes with $M = [1, 3, 5]$ and $N = 32$ when (a) fitnesses are not normalized, (b) fitnesses are normalized by the global maximum and the global minimum of the NM landscapes.

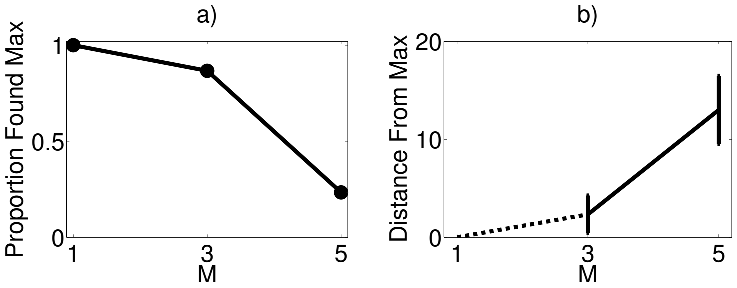


Figure 4.12: Results of search with GA on 32 random NM landscapes with $M = [1, 3, 5]$ and $N = 32$. (a) The proportion of the times the search found the global maximum out of 32 runs. (b) The mean and the standard deviation of the distances between the best solution found by GA and the global maximum, when the global maximum was not found. The dashed line indicates that at $M = 1$ all runs found the global maximum.

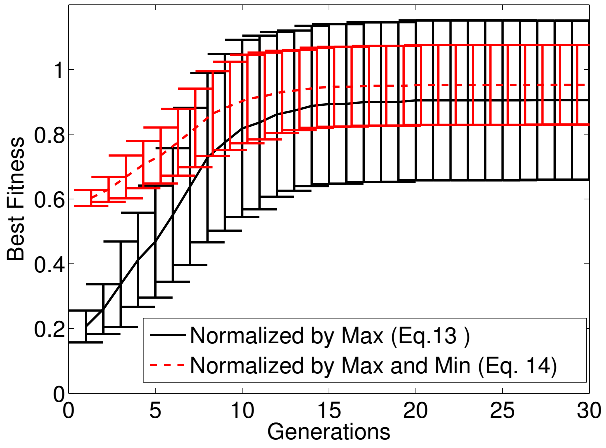


Figure 4.13: The mean and the standard deviation of the best fitnesses found by GA on NM landscapes with $M = 3$ and $N = 32$, when fitnesses are either normalized by only the global maximum (red dashed lines) or both the global maximum and the global minimum (black lines).

The difficulty of GA search on NM landscapes also increases with increasing m and M, by several measures of difficulty. When the maximum order of interactions $M = 2$ and the proportion P of all the possible second-order interactions increases from 0 to 1 in 0.1 increments, our results show that the mean of the best fitnesses found by the GA decreases, although above $P = 0.7$ there is little if any further change in difficulty (Fig. 4.9). We speculate that this might correspond to the periodically asymptotic pattern in the ruggedness noted previously as the maximum number of terms m approaches the maximum possible for a given M (Fig. 4.4a-b). Results are shown for only the first 30 generations, after which no further improvement was observed. Histograms of the Hamming distances between the best solutions found by GA and the global maximum are shown for 32 runs of the GA, for different proportions of the possible second-order interactions (Fig. 4.10). For unimodal landscapes ($M = 1$), the GA found the global optimum in all 32 runs (Fig. 4.10). For more rugged landscapes the global optimum was also found in some runs, and surprisingly the proportion of times it was found increased from $P = 0.2$ to $P = 1$. However, as the ruggedness increased, those runs in which the best individuals were suboptimal generally became stuck farther and farther from the global optimum (note how the distributions become increasingly spread out to the right, as you view the panels in Fig. 4.10 from top to bottom).

When fitnesses are not normalized, a higher maximum order of interactions M results in higher raw fitnesses (Fig. 4.11a). This is due to the fact that summing more interaction terms result in higher ranges of fitness (Fig. 4.6 and Fig. 4.7). However, when fitnesses are properly normalized by Eq. 4.14 to the range [0,1], increasing the maximum order of interactions in the model decreases the values of the best individuals' fitnesses found (Fig. 4.11b), reflecting the fact that GA search becomes more difficult at higher M.

While the proportion of times that GA was able to find the global maximum out of 32 runs decreased as the maximum order of interactions M increased (Fig. 4.12a), the means and standard deviations of the distances between the best solutions found by the GA and the global maximum increased (Fig. 4.12b). Normalizing by Eq. (4.13), rather than Eq. (4.14) exagerates both the apparent relative generational increase in fitnesses in the GA and the variance in fitnesses across different random landscapes with the same m and M (Fig. 4.11). This illustrates how knowing the global minimum can help to assess the relative increase in fitnesses and fairly compare search results on different NM landscapes.

4.6 Discussion

In this work we introduce NM landscapes, which are parametric interaction models that (a) have non-negative coefficients and (b) are well-defined for feature alphabets of any arity (from binary to real-valued), as long as (c) the minimum value in the alphabet is negative with absolute value less than or equal to the positive maximum. This combination of constraints ensures that a global maximum is located at the point where all decision variables have their maximum value, with the optimal value equal to the sum of the model coefficients. By further restricting which combinations of interactions are present, various subsets of NM models also have known location and value of the global minimum (we illustrate two such subsets, which we refer to as Type II and Type III NM landscapes). By using an appropriate non-negative distribution for the coefficients, the resulting NM landscape models have relatively smoothly tunable ruggedness. Epistatic terms are transparently represented in NM landscapes, making it trivial to control or analyze exactly which terms and interactions are present. In the following, we discuss why these various aspects of NM landscapes are valuable, and how they offer advantages over NK landscapes and Walsh polynomials as epistatic benchmark problems.

4.6.1 Value of finely tunable epistasis

Although NK landscapes have been widely used as benchmark problems with varying degrees of epistasis, there are many potential applications that require more fine control over which terms are present or absent.

For example, this study was originally motivated by some of our previous research in comparing search strategies for healthcare improvement [27, 73]. In the context of clinical fitness landscapes, it is not reasonable to assume that all features have only main effects (corresponding to $K = 0$ in NK landscapes) as there are many known interactions between various practices and/or treatments in the real world (e.g., [18, 56]). However, it is also not reasonable to assume that every feature interacts with at least one other feature (corresponding to $K = 1$). Rather, we sought to explore the performance of the different clinical quality improvement strategies (including randomized controlled trials and team quality improvement collaboratives) in more realistic clinical fitness landscapes where all features had main effects but varying numbers of second-order interactions were also present.

Alternatively, in some application domains one may wish to model purely epistatic landscapes in which *no* main effects are present. For example, in complex diseases there may be little if any association between single genes and incidence of disease [81]. Similarly, the electrical grid is explicitly ensured to be stable with respect to the loss of any one component, but interactions between two or more component outages can lead to large cascading failures [26]. For these types of applications, we and others have been seeking algorithms that are capable of detecting purely epistatic interactions (e.g., [25, 26, 132]). To test these algorithms, one must be able to create benchmark landscapes where there are interaction terms but no main effects.

Classic NK landscapes cannot model landscapes between $K = 0$ and $K = 1$, nor can they model landscapes with no main effects or where the strengths of the main effects are smaller than the strengths of interaction terms [15]. In contrast, general interaction models (including NM landscapes) easily allow fine control over exactly which terms are present or absent and one can easily specify different magnitudes of coefficients for different terms. This is also possible using Walsh polynomials, although the notation is not as simple or transparent.

In the experiments shown here, we provide evidence that increasing the number m and/or maximum order M of interactions increases the ruggedness of NM landscapes with coefficients generated using Eq. (4.9) with $\sigma = N$, as measured by number of local peaks and the lag 1 autocorrelation of random walks through the landscapes (Fig. 4.4), and also increases the difficulty of these landscapes by several different measures of search difficulty, including size of the basin of attraction of the global optimum (Fig. 4.8), final normalized best fitnesses found with a GA (Figs. 4.9 and 4.11), distances from the global optimum of sub-optimal final best fitnesses found by a GA (Figs. 4.10 and 4.12a), and proportion of times a GA was able to find the global optimum (Fig. 4.12b).

4.6.2 Value of fitness normalization

Since the range of possible fitness values varies so much between rugged landscapes (as illustrated in Figs. 4.6 and 4.7), it is important to normalize fitnesses to a consistent range if one desires to compare fitness values on different lanscapes (Fig 4.11), or to assess the variability of a search strategy on landscapes with a given m and M (Fig. 4.9). In [27, 73] we used logistic transforms of general parametric interaction models with unknown maxima to model search on clinical fitness lanscapes with varying numbers of second order interactions. While the logistic function successfully bounds the transformed fitnesses to the open interval $(0, 1)$, it also has the side effect of compressing high fitness values to the degree that there is very little difference between the fitnesses of the optimal peak and many suboptimal peaks. This may be a realistic assumption in health care (where there may be many possible combinations of clinical practices that yield good results), but for applications where such compression is not ideal it may be more appropriate to normalize fitnesses to values ≤ 1 using Eq. (4.13), which requires knowing the global maximum, or even better to the closed interval $[0, 1]$ using Eq. (4.14), which also requires knowing the global minimum. NM landscapes enable these types of normalization, as discussed in the following subsections.

4.6.3 Value of knowing the global maximum

Knowing the best possible fitness offers obvious benefits, including: (a) one can terminate a search as soon as the known optimal value is found, potentially saving significant computation time; (b) one can compare methods by assessing the frequency with which the search strategies are able to find the global maximum; (c) one can tell if a stalled search has found the global optimum or is stuck on a local optimum. Knowing the location of the global maximum in feature space offers obvious additional benefits (e.g., [57]) including: (e) one can track the evolving distances of solutions to the global optimum as the search progresses, which could potentially inform ways to improve the search process; (f) one can compare the distances (in feature space) from the best final solution to the global optimum; (g) one can assess the difficulty of the fitness landscape by assessing the correlation of fitness values encountered on a random walk with the distances to the global optimum; (h) one can empirically explore a landscape near the global optimum in order to asses the size and shape of its basin of attraction, and (i) one can normalize fitnesses be ≤ 1 using equation (4.13).

For arbitrary epistatic landscapes (including NK landscapes, general parametric interaction models, and Walsh polynomials) finding the global maximum is NP complete. However, there are restricted subsets of these for which the global maximum is known. For example, in Walsh polynomials one can select an arbitrary point and then adjust the signs of the coefficients to force this to be the global maximum [125]. In NM landscapes both the location and value of the global maximum is trivially known.

4.6.4 Value of knowing the global minimum

While fitnessess can be partially normalized to values ≤ 1 with Eq. 4.13 (as in Fig. 4.9), this can still be misleading, since the range of fitness values has not been properly accounted for. It is thus preferable to normalize to values in the closed interval $[0, 1]$ with Eq. (4.14), as in Fig. 4.11. For example, in Fig. 4.13 we illustrate how both increase in relative fitnesses over time and the variability of fitnesses on different landscapes with the same maximum order M are over-estimated when normalizing by Eq. (4.13), which only requires that the maximum possible fitness value be known, relative to when the data is normalized by Eq. (4.14), which requires that both the maximum and minimum possible fitness values be known.

Finding the global minimum is NP complete in NK landscapes and Walsh polynomials. However, in certain subsets of NM landscapes (e.g., Type II and Type III NM landscapes) the value and location of the global minimum is trivially known, enabling proper normalization of fitnesses.

4.6.5 Value of arbitrary arity of the alphabet

Both NK landscapes and Walsh polynomials are defined for combinatorial problems with binary alphabets [62] [31] [59]. There are also a variety of benchmark problems with tunable difficulty for real-valued alphabets (e.g., [16, 110]). However, in some applications it would be desirable to have one type of model with tunable ruggedness that could be applied to binary alphabets, integer alphabets, real-valued alphabets, or heterogeneous alphabets. For example, in real clinical fitness landscapes, decision variables can have a variety of arities ranging from binary (e.g., whether or not a certain practice is performed) to real-valued (e.g., the amount or duration of application of a particular treatment) [73].

NK landscapes and Walsh polynomials are only defined for binary feature alphabets. NM landscapes are well-defined for alphabets of all arities (including mixed arities); changing the arity does not change the location or value of the global maximum or minimum.

4.6.6 Value of transparency of interactions

Various researchers are working on developing algorithms to try to detect which interactions are present in fitness landscapes and use these inferred interactions to guide the search (e.g., the linkage tree genetic algorithm [128]). Being able to easily control exactly which feature interactions are present and also know the relative strengths of these interactions would facilitate the testing and validation of such approaches, as one could easily see whether the algorithm was properly estimating interaction terms.

NK landscapes offer little control over which interactions are to be included, and once generated it is non-obvious which interaction terms are present or what their coefficients values are (without significant effort [15]). Walsh polynomials present a framework where specific interaction terms can be included or excluded from the model, but the notation can be confusing and obfuscates which terms are present (e.g., see the example in Eq. (4.6)). In NM landscapes, the interaction terms and their coefficients are obvious, since this is how interaction models are defined (e.g., see the example in Eq. (4.5)).

4.7 Summary

We propose a new class of fitness landscapes with tunable degrees of epistasis, referred to as NM landscapes. All NM landscapes have a known global optimum, various subsets of NM landscapes have a known global mimimum (thus permitting proper normalization of fitness values), the ruggedness and search difficulty of NM landscapes can be made to be relatively smoothly tunable, NM landscapes are well-defined on alphabets of any arity, and which epistatic interactions are included in a particular instantiation of an NM landscape is easily controlled or analyzed. In summary, NM landscapes are a simple but powerful class of models that offer several benefits over NK landscapes and Walsh polynomials as benchmark models with tunable epistasis.

4.8 Acknowledgements

This work was funded in part by the NIH Eunice Kennedy Shriver National Institute of Child Health & Human Development award 1R21HD068296.

Bibliography

[1] Roberta Annicchiarico, Ulises Cortés, and Cristina Urdiales. *Agent Technology and E-health*. Springer, 2008.

[2] Ignacio Arnaldo, Iván Contreras, David Millán-Ruiz, J Ignacio Hidalgo, and Natalio Krasnogor. Matching island topologies to problem structure in parallel evolutionary algorithms. *Soft Computing*, pages 1–17, 2013.

[3] Lea R Ayers, Suzanne C Beyea, Marjorie M Godfrey, Doreen C Harper, Eugene C Nelson, and Paul B Batalden. Quality improvement learning collaboratives. *Quality Management in Healthcare*, 14(4):234–247, 2005.

[4] Rosa R Baier, David R Gifford, Gail Patry, Sara M Banks, Therese Rochon, Debra DeSilva, and Joan M Teno. Ameliorating pain in nursing homes: a collaborative quality-improvement project. *Journal of the American Geriatrics Society*, 52(12):1988–1995, 2004.

[5] David W Baker, Steven M Asch, Joan W Keesey, Julie A Brown, Kitty S Chan, Geoffrey Joyce, and Emmett B Keeler. Differences in education, knowledge, self-management activities, and health outcomes for patients with heart failure cared for under the chronic disease model: the improving chronic illness care evaluation. *Journal of cardiac failure*, 11(6):405–413, 2005.

[6] A.S. Banks. Cross-national time-series data archive (cnts) 1815-2007. *Databanks International, Jerusalem, Israel*, 2008.

[7] Alberto Barceló, Elizabeth Cafiero, Melanie de Boer, Alejandro Escobar Mesa, Marcelina García Lopez, Rosa Aurora Jiménez, Agustín Lara Esqueda, José Antonio Martinez, Esperanza Medina Holguin, Micheline Meiners, et al. Using collaborative learning to improve diabetes care and outcomes: The vida project. *Primary care diabetes*, 4(3):145–153, 2010.

[8] Paul B Batalden and Frank Davidoff. What is quality improvement and how can it transform healthcare? *Quality and Safety in Health Care*, 16(1):2–3, 2007.

[9] Jordana T Bell, Nicholas J Timpson, N William Rayner, Eleftheria Zeggini, Timothy M Frayling, Andrew T Hattersley, Andrew P Morris, and Mark I McCarthy. Genome-wide association scan allowing for epistasis in type 2 diabetes. *Annals of human genetics*, 75(1):10–19, 2011.

[10] Rob Benedetti, Barb Flock, Steve Pedersen, et al. Improved clinical outcomes for fee-for-service physician practices participating in a diabetes care collaborative. *Joint Commission Journal on Quality and Patient Safety*, 30(4):187–194, 2004.

[11] I.M. Bernstein, J.D. Horbar, G.J. Badger, A. Ohlsson, A. Golan, et al. Morbidity and mortality among very-low-birth-weight neonates with intrauterine growth restriction. *American journal of obstetrics and gynecology*, 182(1):198–206, 2000.

[12] DM Berwick. Broadening the view of evidence-based medicine. *Quality and Safety in Health Care*, 14(5):315–316, 2005.

[13] Sebastian Bonhoeffer, Colombe Chappey, Neil T Parkin, Jeanette M Whitcomb, and Christos J Petropoulos. Evidence for positive epistasis in hiv-1. *Science*, 306(5701):1547–1550, 2004.

[14] Penny Bundy. Using drama in the counselling process: the moving on project. *Research in drama education*, 11(1):7–18, 2006.

[15] Jeffrey Buzas and Jeffrey Dinitz. An analysis of NK landscapes: Interaction structure, statistical properties and expected number of local optima. *IEEE Transactions on Evolutionary Computation*, in press, DOI10.1109/TEVC.2013.2286352, 2014.

[16] Pilar Caamaño, Abraham Prieto, José Antonio Becerra, Francisco Bellas, and Richard J Duro. Real-valued multimodal fitness landscape characterization for evolution. In *Neural Information Processing. Theory and Algorithms*, pages 567–574. Springer, 2010.

[17] P.R. Cohen and H.J. Levesque. Teamwork. *Nous*, pages 487–512, 1991.

[18] A. Dechartres, I. Boutron, L. Trinquart, et al. Single-center trials show larger treatment effects than multicenter trials: Evidence from a meta-epidemiologic study. *Annals of internal medicine*, 155(1):39, 2011.

[19] D. DeHaas, J. Craig, C. Rickert, P. Haake, K. Stor, and M.J. Eppstein. Feature selection and classification in noisy epistatic problems using a hybrid evolutionary approach. *poster and published extended abstract accepted for Genetic and Evolutionary Computation Conference (GECCO)*, 2007.

[20] OM Dekkers, Erik von Elm, Ale Algra, JA Romijn, and JP Vandenbroucke. How to assess the external validity of therapeutic trials: a conceptual approach. *International journal of epidemiology*, 39(1):89–94, 2010.

[21] J. Denrell and C. Liu. Top performers are not the most impressive when extreme performance indicates unreliability. *Proceedings of the National Academy of Sciences*, 109(24):9331–9336, 2012.

[22] Hui Ding, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, and Eamonn Keogh. Querying and mining of time series data: experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment*, 1(2):1542–1552, 2008.

[23] Lori Ebert, Lisa Amaya-Jackson, Jan M Markiewicz, Cassandra Kisiel, and John A Fairbank. Use of the breakthrough series collaborative to support broad and sustained use of evidence-based trauma treatment for children in community practice settings. *Administration and Policy in Mental Health and Mental Health Services Research*, 39(3):187–199, 2012.

[24] Judith A Effken. Different lenses, improved outcomes: a new approach to the analysis and design of healthcare information systems. *International journal of medical informatics*, 65(1):59–74, 2002.

[25] Margaret J Eppstein and Paul Haake. Very large scale relieff for genome-wide association analysis. In *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 112–119, 2008.

[26] M.J. Eppstein and P.D.H. Hines. A random chemistry; algorithm for identifying collections of multiple contingencies that initiate cascading failure. *IEEE Transactions on Power Systems*, 27(3):1698–1705, 2012.

[27] M.J. Eppstein, J.D. Horbar, J.S. Buzas, and S.A. Kauffman. Searching the clinical fitness landscape. *PLoS ONE*, 7(11):e49901, 2012.

[28] Philippe Esling and Carlos Agon. Time-series data mining. *ACM Computing Surveys (CSUR)*, 45(1):12, 2012.

[29] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.

[30] Usama M Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy. Advances in knowledge discovery and data mining. 1996.

[31] Stephanie Forrest and Melanie Mitchell. The performance of genetic algorithms on walsh polynomials: Some anomalous results and their explanation. In *Proceedings of the 4th International Cinference on Genetic Alogarithms*, pages 182–189. San Mateo, CA: Morgan Kaufmann, 1991.

[32] Sarah W Fraser and Trisha Greenhalgh. Coping with complexity: educating for capability. *BMJ*, 323(7316):799–803, 2001.

[33] Mingxin Gan and Rui Jiang. Constructing a user similarity network to remove adverse influence of popular objects for personalized recommendation. *Expert Systems with Applications*, 2013.

[34] Mingxin Gan and Rui Jiang. Improving accuracy and diversity of personalized recommendation through power law adjustments of user similarities. *Decision Support Systems*, 2013.

[35] Yong Gao and Joseph C. Culberson. An analysis of phase transition in NK landscapes. *Journal of Artificial Intelligence Research*, 17(1):309–332, 2002.

[36] Ilaria Giannoccaro. Complex systems methodologies for behavioural research in operations management: NK fitness landscape. In *Behavioral Issues in Operations Management*, pages 23–47. Springer, 2013.

[37] R.J. Gray. A bayesian analysis of institutional effects in a multicenter cancer clinical trial. *Biometrics*, pages 244–253, 1994.

[38] Jon W Gregersen, Kamil R Kranc, Xiayi Ke, Pia Svendsen, Lars S Madsen, Allan Randrup Thomsen, Lon R Cardon, John I Bell, and Lars Fugger. Functional epistasis on a common mhc haplotype associated with multiple sclerosis. *Nature*, 443(7111):574–577, 2006.

[39] R. Guimera, B. Uzzi, J. Spiro, and L.A.N. Amaral. Team assembly mechanisms determine collaboration network structure and team performance. *Science*, 308(5722):697–702, 2005.

[40] R. Hecht-Nielsen. Counterpropagation networks. *Applied optics*, 26(23):4979–4983, 1987.

[41] L. Hong and S.E. Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the United States of America*, 101(46):16385–16389, 2004.

[42] J.D. Horbar. The vermont oxford network: evidence-based quality improvement for neonatology. *Pediatrics*, 103(Supplement):350, 1999.

[43] J.D. Horbar, G.J. Badger, J.H. Carpenter, A.A. Fanaroff, S. Kilpatrick, M. LaCorte, R. Phibbs, R.F. Soll, et al. Trends in mortality and morbidity for very low birth weight infants, 1991–1999. *Pediatrics*, 110(1):143, 2002.

[44] J.D. Horbar, G.J. Badger, E.M. Lewit, J. Rogowski, P.H. Shiono, et al. Hospital and patient characteristics associated with variation in 28-day mortality rates for very low birth weight infants. *Pediatrics*, 99(2):149, 1997.

[45] J.D. Horbar et al. The vermont-oxford neonatal network: integrating research and clinical practice to improve the quality of medical care. In *Seminars in perinatology*, volume 19, page 124, 1995.

[46] J.D. Horbar and J.F. Lucey. Evaluation of neonatal intensive care technologies. *The Future of Children*, pages 139–161, 1995.

[47] J.D. Horbar, P.E. Plsek, and K. Leahy. Nic/q 2000: establishing habits for improvement in neonatal intensive care units. *Pediatrics*, 111(Supplement):e397, 2003.

[48] J.D. Horbar, P.E. Plsek, J.A. Schriefer, and K. Leahy. Evidence-based quality improvement in neonatal and perinatal medicine: the neonatal intensive care quality improvement collaborative experience. *Pediatrics*, 118(Supplement):S57, 2006.

[49] J.D. Horbar, J. Rogowski, P.E. Plsek, P. Delmore, W.H. Edwards, J. Hocker, A.D. Kantak, P. Lewallen, W. Lewis, E. Lewit, et al. Collaborative quality improvement for neonatal intensive care. *Pediatrics*, 107(1):14–22, 2001.

[50] J.D. Horbar, J. Rogowski, P.E. Plsek, P. Delmore, W.H. Edwards, J. Hocker, A.D. Kantak, P. Lewallen, W. Lewis, E. Lewit, et al. Collaborative quality improvement for neonatal intensive care. *Pediatrics*, 107(1):14–22, 2001.

[51] J.D. Horbar, R.F. Soll, and W.H. Edwards. The vermont oxford network: a community of practice. *Clin Perinatol*, 37(1):29–47, 2010.

[52] J.D. Horbar, R.F. Soll, and W.H. Edwards. The Vermont Oxford Network: A community of practice. *Clinics in perinatology*, 37(1):29, 2010.

[53] Wim Hordijk. Correlation analysis of coupled fitness landscapes. In *Recent Advances in the Theory and Application of Fitness Landscapes*, pages 369–393. Springer, 2014.

[54] R.I. Horwitz, B.H. Singer, R.W. Makuch, and C.M. Viscoli. Can treatment that is helpful on average be harmful to some patients? A study of the conflicting information needs of clinical inquiry and drug regulation. *Journal of Clinical Epidemiology*, 49(4):395–400, 1996.

[55] T.K. Jenssen, W. Kuo, T. Stokke, and E. Hovig. Associations between gene expressions in breast cancer and patient survival. *Human genetics*, 111(4):411–420, 2002.

[56] K. Johnell and I. Klarin. The relationship between number of drugs and potential drug-drug interactions in the elderly: A study of over 600000 elderly patients from the swedish prescribed drug register. *Drug Safety*, 30(10):911–918, 2007.

[57] Terry Jones and Stephanie Forrest. Fitness distance correlation as a measure of problem difficulty for genetic algorithms. In *ICGA*, volume 95, pages 184–192. Citeseer, 1995.

[58] E.T. Juengst, R.A. Settersten, J.R. Fishman, and M.L. McGowan. After the revolution? Ethical and social challenges in personalized genomic medicine. *Personalized Medicine*, 9(4):429–439, 2012.

[59] Leila Kallel, Bart Naudts, and Colin R Reeves. Properties of fitness functions and search landscapes. In *Theoretical aspects of evolutionary computing*, pages 175–206. Springer, 2001.

[60] Stuart Kauffman. *The origins of order: Self organization and selection in evolution*. Oxford University Press, 1993.

[61] Stuart Kauffman and Simon Levin. Towards a general theory of adaptive walks on rugged landscapes. *Journal of theoretical Biology*, 128(1):11–45, 1987.

[62] Stuart A Kauffman and Edward D Weinberger. The NK model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of theoretical biology*, 141(2):211–245, 1989.

[63] Eamonn Keogh and Shruti Kasetty. On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and knowledge discovery*, 7(4):349–371, 2003.

[64] Charles M Kilo. A framework for collaborative improvement: lessons from the institute for healthcare improvement's breakthrough series. *Quality Management in Healthcare*, 6(4):1–14, 1998.

[65] Bobby PC Koeleman, Benedicte Alexandre Lie, Dag Erik Undlien, Frank Dudbridge, Erik Thorsby, Rindert RP De Vries, Francesco Cucca, Bart O Roep, MJ Giphart, and John A Todd. Genotype effects and epistasis in type 1 diabetes and hla-dq trans dimer associations with disease. *Genes and immunity*, 5(5):381–388, 2004.

[66] J.S. Krupa, S. Chatterjee, E. Eldridge, D.M. Rizzo, and M.J. Eppstein. Evolutionary exploratory association discovery: A plug-in hybrid vehicle adoption application. *Submitted to the 21st International GECCO Confference*, 2012.

[67] J.A. LePine, R.F. Piccolo, C.L. Jackson, J.E. Mathieu, and J.R. Saul. A meta-analysis of teamwork processes: Tests of a multidimensional model and relationships with team effectiveness criteria. *Personnel Psychology*, 61(2):273–307, 2008.

[68] Rui Li, Michael TM Emmerich, Jeroen Eggermont, Ernst GP Bovenkamp, Thomas Bäck, Jouke Dijkstra, and Johan HC Reiber. Mixed-integer nk landscapes. In *Parallel Problem Solving from Nature-PPSN IX*, pages 42–51. Springer, 2006.

[69] Rung Tzuo Liaw and Chuan Kang Ting. Effect of model complexity for estimation of distribution algorithm in NK landscapes. In *2013 IEEE Symposium on Foundations of Computational Intelligence (FOCI)*, pages 76–83. IEEE, 2013.

[70] H. Liu, J. Li, and L. Wong. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Informatics Series*, pages 51–60, 2002.

[71] Rita Mangione-Smith, Matthias Schonlau, Kitty S Chan, Joan Keesey, Mayde Rosen, Thomas A Louis, and Emmett Keeler. Measuring the effectiveness of a collaborative for quality improvement in pediatric asthma care: does implementing the chronic care model improve processes and outcomes of care? *Ambulatory Pediatrics*, 5(2):75–82, 2005.

[72] T. Manser. Teamwork and patient safety in dynamic domains of healthcare: A review of the literature. *Acta Anaesthesiologica Scandinavica*, 53(2):143–151, 2008.

[73] Narine Manukyan, Margaret J Eppstein, and Jeffrey D Horbar. Team learning for healthcare quality improvement. *IEEE Access*, 1:545–557, 2013.

[74] Narine Manukyan, Margaret J Eppstein, and Donna M Rizzo. Data-driven cluster reinforcement and visualization in sparsely-matched self-organizing maps. *Neural Networks and Learning Systems, IEEE Transactions on*, 23(5):846–852, 2012.

[75] M.A. Marks, J.E. Mathieu, and S.J. Zaccaro. A temporally based framework and taxonomy of team processes. *Academy of Management Review*, pages 356–376, 2001.

[76] J.A. Martin, K.D. Kochanek, D.M. Strobino, B. Guyer, and M.F. MacDorman. Annual summary of vital statistics—2003. *Pediatrics*, 115(3):619, 2005.

[77] Klim McPherson, John E Wennberg, Ole B Hovind, Peter Clifford, et al. Small-area variations in the use of common surgical procedures: An international comparison of New England, England, and Norway. *The New England journal of medicine*, 307(21):1310, 1982.

[78] Brian S Mittman. Creating the evidence base for quality improvement collaboratives. *Annals of internal medicine*, 140(11):897–901, 2004.

[79] Naoki Miyagawa, Hiroshi Teramoto, Chun-Biu Li, and Tamiki Komatsuzaki. Decomposability of multivariate interactions. *Complex Systems*, 20(2):165, 2011.

[80] Douglas C Montgomery, Douglas C Montgomery, and Douglas C Montgomery. *Design and analysis of experiments*, volume 7. Wiley New York, 1984.

[81] Jason H Moore. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human heredity*, 56(1-3):73–82, 2003.

[82] Alberto Moraglio and Julian Togelius. Geometric differential evolution. In *Proceedings of the 11th Annual conference on Genetic and evolutionary computation*, pages 1705–1712. ACM, 2009.

[83] L.S. Morales, D. Staiger, J.D. Horbar, J. Carpenter, M. Kenny, J. Geppert, and J. Rogowski. Mortality among very low birthweight infants in hospitals serving minority populations. *American journal of public health*, 95(12):2206, 2005.

[84] Erum Nadeem, S Serene Olin, Laura Campbell Hill, Kimberly Eaton Hoagwood, and Sarah McCue Horwitz. Understanding the components of quality improvement collaboratives: A systematic literature review. *Milbank Quarterly*, 91(2):354–394, 2013.

[85] P.J. Newton, EJ Halcomb, PM Davidson, and A.R. Denniss. Barriers and facilitators to the implementation of the collaborative method: Reflections from a single site. *Quality and Safety in Health Care*, 16(6):409–414, 2007.

[86] I.S. Oh, J.S. Lee, and B.R. Moon. Hybrid genetic algorithms for feature selection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(11):1424–1437, 2004.

[87] World Health Organization et al. Who patient safety curriculum guide for medical schools. 2009.

[88] J. Øvretveit, P. Bate, P. Cleary, S. Cretin, D. Gustafson, K. McInnes, H. McLeod, T. Molfenter, P. Plsek, G. Robert, et al. Quality collaboratives: lessons from research. *Quality and safety in health care*, 11(4):345–351, 2002.

[89] J Øvretveit, Paul Bate, Paul Cleary, Shan Cretin, D Gustafson, K McInnes, H McLeod, T Molfenter, P Plsek, Glenn Robert, et al. Quality collaboratives: lessons from research. *Quality and safety in health care*, 11(4):345–351, 2002.

[90] Ray Pawson and Nick Tilley. *Realistic evaluation*. Sage, 1997.

[91] N.R. Payne, M.J. Finkelstein, M. Liu, J.W. Kaempf, P.J. Sharek, and S. Olsen. Nicu practices and outcomes associated with 9 years of quality improvement collaboratives. *Pediatrics*, 125(3):437–446, 2010.

[92] Martin Pelikan. Analysis of estimation of distribution algorithms and genetic algorithms on NK landscapes. In *Proceedings of the 10th annual conference on Genetic and evolutionary computation*, pages 1033–1040. ACM, 2008.

[93] F. Pernkopf and P. O'Leary. Feature selection for classification using genetic algorithms with a novel encoding. In *Computer Analysis of Images and Patterns*, pages 161–168. Springer, 2001.

[94] Charles Perrow. *Normal Accidents: Living with High Risk Technologies (Updated)*. Princeton University Press, 2011.

[95] Paul E Plsek. Collaborating across organizational boundaries to improve the quality of care. *American journal of infection control*, 25(2):85–95, 1997.

[96] Paul E Plsek and Trisha Greenhalgh. The challenge of complexity in health care. *Bmj*, 323(7313):625–628, 2001.

[97] Paul E Plsek and Tim Wilson. Complexity, leadership, and management in healthcare organisations. *Bmj*, 323(7315):746–749, 2001.

[98] P.E. Plsek. Collaborating across organizational boundaries to improve the quality of care. *American journal of infection control*, 25(2):85–95, 1997.

[99] B.C. Poulton and M.A. West. Effective multidisciplinary teamwork in primary health care. *Journal of Advanced Nursing*, 18(6):918–925, 2008.

[100] M.L. Raymer, W.F. Punch, E.D. Goodman, L.A. Kuhn, and A.K. Jain. Dimensionality reduction using genetic algorithms. *Evolutionary Computation, IEEE Transactions on*, 4(2):164–171, 2000.

[101] Colin Reeves and Christine Wright. An experimental design perspective on genetic algorithms. In *Foundations of Genetic Algorithms 3*, 1995.

[102] Colin R Reeves. Experiments with tuneable fitness landscapes. In *Parallel Problem Solving from Nature PPSN VI*, pages 139–148. Springer, 2000.

[103] Colin R Reeves and Christine C Wright. Epistasis in genetic algorithms: An experimental design perspective. In *Proceedings of the 6th International Conference on Genetic Algorithms*, pages 217–224. Morgan Kaufmann Publishers Inc., 1995.

[104] Ian Reid. Complexity science: Let them eat complexity: the emperor's new toolkit. *BMJ: British Medical Journal*, 324(7330):171, 2002.

[105] D.N. Reshef, Y.A. Reshef, H.K. Finucane, S.R. Grossman, G. McVean, P.J. Turnbaugh, E.S. Lander, M. Mitzenmacher, and P.C. Sabeti. Detecting novel associations in large data sets. *science*, 334(6062):1518–1524, 2011.

[106] A.D. Rodrigues. *Drug-drug interactions*. Informa Healthcare, New York, NY, 2008.

[107] J.A. Rogowski, J.D. Horbar, P.E. Plsek, L.S. Baker, J. Deterding, W.H. Edwards, J. Hocker, A.D. Kantak, P. Lewallen, W. Lewis, et al. Economic implications of neonatal intensive care unit collaborative quality improvement. *Pediatrics*, 107(1):23, 2001.

[108] J.A. Rogowski, J.D. Horbar, D.O. Staiger, M. Kenny, J. Carpenter, and J. Geppert. Indirect vs direct hospital quality indicators for very low-birth-weight infants. *JAMA: the journal of the American Medical Association*, 291(2):202, 2004.

[109] J.A. Rogowski, D.O. Staiger, and J.D. Horbar. Variations in the quality of care for very-low-birthweight infants: implications for policy. *Health Affairs*, 23(5):88–97, 2004.

[110] Jani Rönkkönen, Xiaodong Li, Ville Kyrki, and Jouni Lampinen. A framework for generating tunable test functions for multimodal optimization. *Soft Computing*, 15(9):1689–1706, 2011.

[111] Peter M Rothwell. External validity of randomised controlled trials:to whom do the results of this trial apply?. *The Lancet*, 365(9453):82–93, 2005.

[112] William Rowe, Mark Platt, David C Wedge, Philip J Day, Douglas B Kell, and Joshua Knowles. Analysis of a complete dna–protein affinity landscape. *Journal of The Royal Society Interface*, 7(44):397–408, 2010.

[113] Bill Runciman and Merrilyn Walton. *Safety and ethics in healthcare: a guide to getting it right*. Ashgate Publishing, Ltd., 2007.

[114] E. Salas, N.J. Cooke, and M.A. Rosen. On teams, teamwork, and team performance: Discoveries and developments. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 50(3):540–547, 2008.

[115] T. Sandmann and M. Boutros. Screens, maps & networks: From genome sequences to personalized medicine. *Current Opinion in Genetics & Development*, 22:36–44, 2012.

[116] Elad Schneidman, Susanne Still, Michael J Berry, William Bialek, et al. Network information and connected correlations. *Physical review letters*, 91(23):238701, 2003.

[117] Matthias Schonlau, Rita Mangione-Smith, Kitty S Chan, Joan Keesey, Mayde Rosen, Thomas A Louis, Shin-Yi Wu, and Emmett Keeler. Evaluation of a quality improvement collaborative in asthma care: does it improve processes and outcomes of care? *The Annals of Family Medicine*, 3(3):200–208, 2005.

[118] L.M.T. Schouten, R.P.T.M. Grol, and M.E.J.L. Hulscher. Factors influencing success in quality-improvement collaboratives: Development and psychometric testing of an instrument. *Implementation Science*, 5(1):1–9, 2010.

[119] L.M.T. Schouten, M.E.J.L. Hulscher, J.J.E. Everdingen, R. Huijsman, and R.P.T.M. Grol. Evidence for the impact of quality improvement collaboratives: Systematic review. *Bmj*, 336(7659):1491–1494, 2008.

[120] Loes MT Schouten, MEJL Hulscher, Jannes JE Van Everdingen, Robbert Huijsman, Louis W Niessen, and RPTM Grol. Short-and long-term effects of a quality improvement collaborative on diabetes management. *Implement Sci*, 5:94, 2010.

[121] Paul G Shekelle, Peter J Pronovost, Robert M Wachter, Stephanie L Taylor, Sydney M Dy, Robbie Foy, Susanne Hempel, Kathryn M McDonald, John Ovretveit, Lisa V Rubenstein, et al. Advancing the science of patient safety. *Annals of Internal Medicine*, 154(10):693–696, 2011.

[122] Stephen M Shortell, Jill A Marsteller, Michael Lin, Marjorie L Pearson, Shin-Yi Wu, Peter Mendel, Shan Cretin, and Mayde Rosen. The role of perceived team effectiveness in improving chronic illness care. *Medical care*, 42(11):1040–1048, 2004.

[123] Leif I Solberg. If youve seen one quality improvement collaborative. *The Annals of Family Medicine*, 3(3):198–199, 2005.

[124] Kenneth Tan, Gordon Baxter, Simon Newell, Steve Smye, Peter Dear, Keith Brownlee, and Jonathan Darling. Knowledge elicitation for validation of a neonatal ventilation expert system utilising modified delphi and focus group techniques. *International journal of human-computer studies*, 68(6):344–354, 2010.

[125] Reiko Tanese. *Distributed genetic algorithms for Function Optimization*. PhD thesis, The University of Michigan, Ann Arbor, MI, 1989.

[126] S.L. Taylor, S. Dy, R. Foy, et al. What context features might be important determinants of the effectiveness of patient safety practice interventions? *BMJ Quality & Safety*, 20(7):611–617, 2011.

[127] Dirk Thierens. The linkage tree genetic algorithm. In *Parallel Problem Solving from Nature, PPSN XI*, pages 264–273. Springer, 2010.

[128] Dirk Thierens and Peter AN Bosman. Hierarchical problem solving with the linkage tree genetic algorithm. In *Proceeding of the fifteenth annual conference on Genetic and evolutionary computation conference*, pages 877–884. ACM, 2013.

[129] Nicholas Tomko, Inman Harvey, and Andrew Philippides. Unconstrain the population: The benefits of horizontal gene transfer in genetic algorithms. In *SmartData*, pages 117–127. Springer, 2013.

[130] Shaun Treweek and Merrick Zwarenstein. Making trials matter: pragmatic and explanatory trials and the problem of applicability. *Trials*, 10(37):9, 2009.

[131] M.E. Turner. *Groups at work: Theory and research*. Lawrence Erlbaum, Hillsdale, NJ, 2000.

[132] Ryan J Urbanowicz and Jason H Moore. The application of michigan-style learning classifiersystems to address genetic heterogeneity and epistasisin association studies. In *Proceedings of the 12th annual conference on Genetic and evolutionary computation*, pages 195–202. ACM, 2010.

[133] G. Vaidyanathan. Redefining clinical trials: The age of personalized medicine. *Cell*, 148(6):1079–1080, 2012.

[134] Vesselin K Vassilev, Terence C Fogarty, and Julian F Miller. Information characteristics and the structure of landscapes. *Evolutionary Computation*, 8(1):31–60, 2000.

[135] Edward Weinberger. Correlated and uncorrelated fitness landscapes and how to tell the difference. *Biological cybernetics*, 63(5):325–336, 1990.

[136] John Wennberg and Alan Gittelsohn. Small area variations in health care delivery: A population-based health information system can guide planning and regulatory decision-making. *Science*, 182(4117):1102–1108, 1973.

[137] John E Wennberg. *Tracking Medicine: A Researcher's Quest to Understand Health Care*. Oxford University Press, USA, 2010.

[138] M.A. West. *Effective teamwork: Practical lessons from organizational research*. Blackwell Publishing, Oxford, 2012.

[139] Tim Wilson, Donald M Berwick, and Paul D Cleary. What do collaborative improvement projects do? experience from seven countries. *Joint Commission Journal on Quality and Patient Safety*, 29(2):85–93, 2003.

[140] Tim Wilson, Tim Holt, and Trisha Greenhalgh. Complexity and clinical care. *Bmj*, 323(7314):685–688, 2001.

[141] David D Woods, Leila J Johannesen, Richard I Cook, and Nadine B Sarter. Behind human error: Cognitive systems, computers and hindsight. Technical report, DTIC Document, 1994.

[142] David D Woods, Emily S Patterson, and Richard I Cook. Behind human error: taming complexity to improve patient safety. *Handbook of Human Factors and Ergonomics in Health Care and Patient Safety. London: Lawrence Erlbaum*, pages 459–76, 2007.

[143] Michael Wooldridge. *An introduction to multiagent systems*. John Wiley & Sons, 2009.

[144] Alden H Wright, Richard K Thompson, and Jian Zhang. The computational complexity of NK fitness functions. *IEEE Transactions on Evolutionary Computation*, 4(4):373–379, 2000.

[145] S. Wright. The roles of mutation, inbreeding, crossbreeding and selection in evolution. In *Proceedings of the sixth international congress on genetics*, volume 1, pages 356–366, 1932.

[146] Paul C Young, Gordon B Glade, Gregory J Stoddard, and Chuck Norlin. Evaluation of a learning collaborative to improve the delivery of preventive services by pediatric practices. *Pediatrics*, 117(5):1469–1476, 2006.

[147] Z.J. Yu, F. Haghighat, B. Fung, and L. Zhou. A novel methodology for knowledge discovery through mining associations between building operational data. *Energy and Buildings*, 2011.

[148] J.A.F. Zupancic, D.K. Richardson, J.D. Horbar, J.H. Carpenter, S.K. Lee, G.J. Escobar, et al. Revalidation of the score for neonatal acute physiology in the vermont oxford network. *Pediatrics*, 119(1):e156–e163, 2007.

Chapter 5

Concluding Remarks

5.1 Summary of Main Findings of Dissertation

This dissertation presents several methods for studying healthcare quality improvement initiatives using computational methods. In the first part of the dissertation we present a genetic algorithm for co-evolving four important aspects of exploratory multivariate time-series clinical data on inter-hospital interactions aimed at quality improvement in healthcare outcomes: (i) a subset of features based on inter-hospital interactions aimed at healthcare improvement to be used as input into some sort of statistical predictor, (ii) which health outcome attribute we can best predict from these input features, (iii) a dividing year that partitions the time-series, and (iv) a time lag to be added to the dividing year that predicts the delay between inter-hospital interactions designed to improve healthcare and subsequent observable changes in health outcomes. While this method correctly inferred interactions on synthetic data, the complexity and incompleteness of the real hospital dataset available to us made it difficult to infer much about the real system, although we did find that participation in QICs was associated with change in health outcomes.

In the second part of the dissertation we present an extension of the agent-based model (ABM) introduced in [27], which permits exploration of various theoretical questions about quality improvement collaboratives (QICs). We use this ABM to compare how different details of team formation affect improvements in patient outcomes. In summary, we found that teams with higher within-team similarity are able to improve performance more quickly than diverse teams, are less sensitive to a variety of factors, and larger teams of similar agents generally perform better than smaller teams. Notably, the advantage of within-team similarity increases with the complexity of the fitness landscape and with the level of noise in fitness evaluation. Based on these results, we propose a new virtual collaboration system that would allow hospitals to receive personalized recommendations about practices for potentially high impact improvement in patient outcomes. This system would provide a mechanism for protecting the privacy of hospital and patient data, while facilitating learning from a common knowledge bank and accounting for differences in local contexts.

To use agent-based modeling for testing different hypotheses on healthcare improvement, it is important to have benchmark landscapes that can properly simulate various characteristics of clinical fitness landscapes. Unfortunately currently existing state-of-the art benchmark landscapes (including NK landscapes and Walsh polynomials) have many limitations. In the third part of this dissertation we thus introduce a new set of benchmark landscapes that we call NM landscapes. NM landscapes have relatively smoothly tunable ruggedness, known global optima, a transparent representation of feature interactions, and are well-defined for both discrete and real-valued features. Subsets of NM landscapes also have known global minima, which permits proper normalization. In conclusion NM landscapes are a simple but powerful class of models that offer several benefits over NK

landscapes and Walsh polynomials as benchmark models with tunable epistasis, making them well suited for modeling clinical fitness landscapes.

5.2 Future Work

While organizations like the Vermont Oxford Network (VON) have been collecting large amounts of data for over two decades, this data collection was not done with the aim of studying how social interactions affect healthcare improvement and thus are insufficient to measure the healthcare implications of social interactions such as QICs. In the future, data collection could focus on creating a richer data environment for analysis of how inter-hospital quality improvement interactions affect healthcare outcomes. VON annual member surveys could be restructured to collect data on the types of day to day clinical practices and treatments that are often the focus of study in QICs. Improvements could also include automatic collection of data on how much practitioners interact with each other via email, phone calls and listserves. In addition, VON members participating in QICs could be more thoroughly surveyed about details on their social interactions designed to improve healthcare.

Many scientists have tried to detect multivariate interactions in complex systems, given a sample of system states [79] [116]. In [116] the authors use the measure of connected information, which is based on Shannon information entropy measure. Others tried to use Shannon entropy from a sample of the possible network states to identify linkages among features to help guide evolutionary search in NK landscapes [127]. There is also much research focused on finding epistasis in real clinical landscapes [13] [38] [65] [9]. Knowing those interactions could potentially improve healthcare outcomes. However, without a clear understand of which interactions are actually present in complex benchmark landscapes used to test these methods, it is difficult to assess how accurately such methods can infer correct interaction terms. In future work we propose to search for the epistatic interactions among multiple practices and patient outcomes in the VON database of member hospitals using the clustering algorithm in [74] applied to the principal components of healthcare practices and outcomes. NM benchmark landscapes will be used to provide an appropriate means for testing and validating our approach, as well as the approach in [127], to finding epistatic interactions.

Bibliography

[1] Roberta Annicchiarico, Ulises Cortés, and Cristina Urdiales. *Agent Technology and E-health*. Springer, 2008.

[2] Ignacio Arnaldo, Iván Contreras, David Millán-Ruiz, J Ignacio Hidalgo, and Natalio Krasnogor. Matching island topologies to problem structure in parallel evolutionary algorithms. *Soft Computing*, pages 1–17, 2013.

[3] Lea R Ayers, Suzanne C Beyea, Marjorie M Godfrey, Doreen C Harper, Eugene C Nelson, and Paul B Batalden. Quality improvement learning collaboratives. *Quality Management in Healthcare*, 14(4):234–247, 2005.

[4] Rosa R Baier, David R Gifford, Gail Patry, Sara M Banks, Therese Rochon, Debra DeSilva, and Joan M Teno. Ameliorating pain in nursing homes: a collaborative quality-improvement project. *Journal of the American Geriatrics Society*, 52(12):1988–1995, 2004.

[5] David W Baker, Steven M Asch, Joan W Keesey, Julie A Brown, Kitty S Chan, Geoffrey Joyce, and Emmett B Keeler. Differences in education, knowledge, self-management activities, and health outcomes for patients with heart failure cared for under the chronic disease model: the improving chronic illness care evaluation. *Journal of cardiac failure*, 11(6):405–413, 2005.

[6] A.S. Banks. Cross-national time-series data archive (cnts) 1815-2007. *Databanks International, Jerusalem, Israel*, 2008.

[7] Alberto Barceló, Elizabeth Cafiero, Melanie de Boer, Alejandro Escobar Mesa, Marcelina García Lopez, Rosa Aurora Jiménez, Agustín Lara Esqueda, José Antonio Martinez, Esperanza Medina Holguin, Micheline Meiners, et al. Using collaborative learning to improve diabetes care and outcomes: The vida project. *Primary care diabetes*, 4(3):145–153, 2010.

[8] Paul B Batalden and Frank Davidoff. What is quality improvement and how can it transform healthcare? *Quality and Safety in Health Care*, 16(1):2–3, 2007.

[9] Jordana T Bell, Nicholas J Timpson, N William Rayner, Eleftheria Zeggini, Timothy M Frayling, Andrew T Hattersley, Andrew P Morris, and Mark I McCarthy. Genome-wide association scan allowing for epistasis in type 2 diabetes. *Annals of human genetics*, 75(1):10–19, 2011.

[10] Rob Benedetti, Barb Flock, Steve Pedersen, et al. Improved clinical outcomes for fee-for-service physician practices participating in a diabetes care collaborative. *Joint Commission Journal on Quality and Patient Safety*, 30(4):187–194, 2004.

[11] I.M. Bernstein, J.D. Horbar, G.J. Badger, A. Ohlsson, A. Golan, et al. Morbidity and mortality among very-low-birth-weight neonates with intrauterine growth restriction. *American journal of obstetrics and gynecology*, 182(1):198–206, 2000.

[12] DM Berwick. Broadening the view of evidence-based medicine. *Quality and Safety in Health Care*, 14(5):315–316, 2005.

[13] Sebastian Bonhoeffer, Colombe Chappey, Neil T Parkin, Jeanette M Whitcomb, and Christos J Petropoulos. Evidence for positive epistasis in hiv-1. *Science*, 306(5701):1547–1550, 2004.

[14] Penny Bundy. Using drama in the counselling process: the moving on project. *Research in drama education*, 11(1):7–18, 2006.

[15] Jeffrey Buzas and Jeffrey Dinitz. An analysis of NK landscapes: Interaction structure, statistical properties and expected number of local optima. *IEEE Transactions on Evolutionary Computation*, in press, DOI10.1109/TEVC.2013.2286352, 2014.

[16] Pilar Caamaño, Abraham Prieto, José Antonio Becerra, Francisco Bellas, and Richard J Duro. Real-valued multimodal fitness landscape characterization for evolution. In *Neural Information Processing. Theory and Algorithms*, pages 567–574. Springer, 2010.

[17] P.R. Cohen and H.J. Levesque. Teamwork. *Nous*, pages 487–512, 1991.

[18] A. Dechartres, I. Boutron, L. Trinquart, et al. Single-center trials show larger treatment effects than multicenter trials: Evidence from a meta-epidemiologic study. *Annals of internal medicine*, 155(1):39, 2011.

[19] D. DeHaas, J. Craig, C. Rickert, P. Haake, K. Stor, and M.J. Eppstein. Feature selection and classification in noisy epistatic problems using a hybrid evolutionary approach. *poster and published extended abstract accepted for Genetic and Evolutionary Computation Conference (GECCO)*, 2007.

[20] OM Dekkers, Erik von Elm, Ale Algra, JA Romijn, and JP Vandenbroucke. How to assess the external validity of therapeutic trials: a conceptual approach. *International journal of epidemiology*, 39(1):89–94, 2010.

[21] J. Denrell and C. Liu. Top performers are not the most impressive when extreme performance indicates unreliability. *Proceedings of the National Academy of Sciences*, 109(24):9331–9336, 2012.

[22] Hui Ding, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, and Eamonn Keogh. Querying and mining of time series data: experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment*, 1(2):1542–1552, 2008.

[23] Lori Ebert, Lisa Amaya-Jackson, Jan M Markiewicz, Cassandra Kisiel, and John A Fairbank. Use of the breakthrough series collaborative to support broad and sustained use of evidence-based trauma treatment for children in community practice settings. *Administration and Policy in Mental Health and Mental Health Services Research*, 39(3):187–199, 2012.

[24] Judith A Effken. Different lenses, improved outcomes: a new approach to the analysis and design of healthcare information systems. *International journal of medical informatics*, 65(1):59–74, 2002.

[25] Margaret J Eppstein and Paul Haake. Very large scale relieff for genome-wide association analysis. In *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 112–119, 2008.

[26] M.J. Eppstein and P.D.H. Hines. A random chemistry; algorithm for identifying collections of multiple contingencies that initiate cascading failure. *IEEE Transactions on Power Systems*, 27(3):1698–1705, 2012.

[27] M.J. Eppstein, J.D. Horbar, J.S. Buzas, and S.A. Kauffman. Searching the clinical fitness landscape. *PLoS ONE*, 7(11):e49901, 2012.

[28] Philippe Esling and Carlos Agon. Time-series data mining. *ACM Computing Surveys (CSUR)*, 45(1):12, 2012.

[29] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.

[30] Usama M Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy. Advances in knowledge discovery and data mining. 1996.

[31] Stephanie Forrest and Melanie Mitchell. The performance of genetic algorithms on walsh polynomials: Some anomalous results and their explanation. In *Proceedings of the 4th International Cinference on Genetic Alogarithms*, pages 182–189. San Mateo, CA: Morgan Kaufmann, 1991.

[32] Sarah W Fraser and Trisha Greenhalgh. Coping with complexity: educating for capability. *BMJ*, 323(7316):799–803, 2001.

[33] Mingxin Gan and Rui Jiang. Constructing a user similarity network to remove adverse influence of popular objects for personalized recommendation. *Expert Systems with Applications*, 2013.

[34] Mingxin Gan and Rui Jiang. Improving accuracy and diversity of personalized recommendation through power law adjustments of user similarities. *Decision Support Systems*, 2013.

[35] Yong Gao and Joseph C. Culberson. An analysis of phase transition in NK landscapes. *Journal of Artificial Intelligence Research*, 17(1):309–332, 2002.

[36] Ilaria Giannoccaro. Complex systems methodologies for behavioural research in operations management: NK fitness landscape. In *Behavioral Issues in Operations Management*, pages 23–47. Springer, 2013.

[37] R.J. Gray. A bayesian analysis of institutional effects in a multicenter cancer clinical trial. *Biometrics*, pages 244–253, 1994.

[38] Jon W Gregersen, Kamil R Kranc, Xiayi Ke, Pia Svendsen, Lars S Madsen, Allan Randrup Thomsen, Lon R Cardon, John I Bell, and Lars Fugger. Functional epistasis on a common mhc haplotype associated with multiple sclerosis. *Nature*, 443(7111):574–577, 2006.

[39] R. Guimera, B. Uzzi, J. Spiro, and L.A.N. Amaral. Team assembly mechanisms determine collaboration network structure and team performance. *Science*, 308(5722):697–702, 2005.

[40] R. Hecht-Nielsen. Counterpropagation networks. *Applied optics*, 26(23):4979–4983, 1987.

[41] L. Hong and S.E. Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the United States of America*, 101(46):16385–16389, 2004.

[42] J.D. Horbar. The vermont oxford network: evidence-based quality improvement for neonatology. *Pediatrics*, 103(Supplement):350, 1999.

[43] J.D. Horbar, G.J. Badger, J.H. Carpenter, A.A. Fanaroff, S. Kilpatrick, M. LaCorte, R. Phibbs, R.F. Soll, et al. Trends in mortality and morbidity for very low birth weight infants, 1991–1999. *Pediatrics*, 110(1):143, 2002.

[44] J.D. Horbar, G.J. Badger, E.M. Lewit, J. Rogowski, P.H. Shiono, et al. Hospital and patient characteristics associated with variation in 28-day mortality rates for very low birth weight infants. *Pediatrics*, 99(2):149, 1997.

[45] J.D. Horbar et al. The vermont-oxford neonatal network: integrating research and clinical practice to improve the quality of medical care. In *Seminars in perinatology*, volume 19, page 124, 1995.

[46] J.D. Horbar and J.F. Lucey. Evaluation of neonatal intensive care technologies. *The Future of Children*, pages 139–161, 1995.

[47] J.D. Horbar, P.E. Plsek, and K. Leahy. Nic/q 2000: establishing habits for improvement in neonatal intensive care units. *Pediatrics*, 111(Supplement):e397, 2003.

[48] J.D. Horbar, P.E. Plsek, J.A. Schriefer, and K. Leahy. Evidence-based quality improvement in neonatal and perinatal medicine: the neonatal intensive care quality improvement collaborative experience. *Pediatrics*, 118(Supplement):S57, 2006.

[49] J.D. Horbar, J. Rogowski, P.E. Plsek, P. Delmore, W.H. Edwards, J. Hocker, A.D. Kantak, P. Lewallen, W. Lewis, E. Lewit, et al. Collaborative quality improvement for neonatal intensive care. *Pediatrics*, 107(1):14–22, 2001.

[50] J.D. Horbar, J. Rogowski, P.E. Plsek, P. Delmore, W.H. Edwards, J. Hocker, A.D. Kantak, P. Lewallen, W. Lewis, E. Lewit, et al. Collaborative quality improvement for neonatal intensive care. *Pediatrics*, 107(1):14–22, 2001.

[51] J.D. Horbar, R.F. Soll, and W.H. Edwards. The vermont oxford network: a community of practice. *Clin Perinatol*, 37(1):29–47, 2010.

[52] J.D. Horbar, R.F. Soll, and W.H. Edwards. The Vermont Oxford Network: A community of practice. *Clinics in perinatology*, 37(1):29, 2010.

[53] Wim Hordijk. Correlation analysis of coupled fitness landscapes. In *Recent Advances in the Theory and Application of Fitness Landscapes*, pages 369–393. Springer, 2014.

[54] R.I. Horwitz, B.H. Singer, R.W. Makuch, and C.M. Viscoli. Can treatment that is helpful on average be harmful to some patients? A study of the conflicting information needs of clinical inquiry and drug regulation. *Journal of Clinical Epidemiology*, 49(4):395–400, 1996.

[55] T.K. Jenssen, W. Kuo, T. Stokke, and E. Hovig. Associations between gene expressions in breast cancer and patient survival. *Human genetics*, 111(4):411–420, 2002.

[56] K. Johnell and I. Klarin. The relationship between number of drugs and potential drug-drug interactions in the elderly: A study of over 600000 elderly patients from the swedish prescribed drug register. *Drug Safety*, 30(10):911–918, 2007.

[57] Terry Jones and Stephanie Forrest. Fitness distance correlation as a measure of problem difficulty for genetic algorithms. In *ICGA*, volume 95, pages 184–192. Citeseer, 1995.

[58] E.T. Juengst, R.A. Settersten, J.R. Fishman, and M.L. McGowan. After the revolution? Ethical and social challenges in personalized genomic medicine. *Personalized Medicine*, 9(4):429–439, 2012.

[59] Leila Kallel, Bart Naudts, and Colin R Reeves. Properties of fitness functions and search landscapes. In *Theoretical aspects of evolutionary computing*, pages 175–206. Springer, 2001.

[60] Stuart Kauffman. *The origins of order: Self organization and selection in evolution*. Oxford University Press, 1993.

[61] Stuart Kauffman and Simon Levin. Towards a general theory of adaptive walks on rugged landscapes. *Journal of theoretical Biology*, 128(1):11–45, 1987.

[62] Stuart A Kauffman and Edward D Weinberger. The NK model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of theoretical biology*, 141(2):211–245, 1989.

[63] Eamonn Keogh and Shruti Kasetty. On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and knowledge discovery*, 7(4):349–371, 2003.

[64] Charles M Kilo. A framework for collaborative improvement: lessons from the institute for healthcare improvement's breakthrough series. *Quality Management in Healthcare*, 6(4):1–14, 1998.

[65] Bobby PC Koeleman, Benedicte Alexandre Lie, Dag Erik Undlien, Frank Dudbridge, Erik Thorsby, Rindert RP De Vries, Francesco Cucca, Bart O Roep, MJ Giphart, and John A Todd. Genotype effects and epistasis in type 1 diabetes and hla-dq trans dimer associations with disease. *Genes and immunity*, 5(5):381–388, 2004.

[66] J.S. Krupa, S. Chatterjee, E. Eldridge, D.M. Rizzo, and M.J. Eppstein. Evolutionary exploratory association discovery: A plug-in hybrid vehicle adoption application. *Submitted to the 21st International GECCO Confference*, 2012.

[67] J.A. LePine, R.F. Piccolo, C.L. Jackson, J.E. Mathieu, and J.R. Saul. A meta-analysis of teamwork processes: Tests of a multidimensional model and relationships with team effectiveness criteria. *Personnel Psychology*, 61(2):273–307, 2008.

[68] Rui Li, Michael TM Emmerich, Jeroen Eggermont, Ernst GP Bovenkamp, Thomas Bäck, Jouke Dijkstra, and Johan HC Reiber. Mixed-integer nk landscapes. In *Parallel Problem Solving from Nature-PPSN IX*, pages 42–51. Springer, 2006.

[69] Rung Tzuo Liaw and Chuan Kang Ting. Effect of model complexity for estimation of distribution algorithm in NK landscapes. In *2013 IEEE Symposium on Foundations of Computational Intelligence (FOCI)*, pages 76–83. IEEE, 2013.

[70] H. Liu, J. Li, and L. Wong. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Informatics Series*, pages 51–60, 2002.

[71] Rita Mangione-Smith, Matthias Schonlau, Kitty S Chan, Joan Keesey, Mayde Rosen, Thomas A Louis, and Emmett Keeler. Measuring the effectiveness of a collaborative for quality improvement in pediatric asthma care: does implementing the chronic care model improve processes and outcomes of care? *Ambulatory Pediatrics*, 5(2):75–82, 2005.

[72] T. Manser. Teamwork and patient safety in dynamic domains of healthcare: A review of the literature. *Acta Anaesthesiologica Scandinavica*, 53(2):143–151, 2008.

[73] Narine Manukyan, Margaret J Eppstein, and Jeffrey D Horbar. Team learning for healthcare quality improvement. *IEEE Access*, 1:545–557, 2013.

[74] Narine Manukyan, Margaret J Eppstein, and Donna M Rizzo. Data-driven cluster reinforcement and visualization in sparsely-matched self-organizing maps. *Neural Networks and Learning Systems, IEEE Transactions on*, 23(5):846–852, 2012.

[75] M.A. Marks, J.E. Mathieu, and S.J. Zaccaro. A temporally based framework and taxonomy of team processes. *Academy of Management Review*, pages 356–376, 2001.

[76] J.A. Martin, K.D. Kochanek, D.M. Strobino, B. Guyer, and M.F. MacDorman. Annual summary of vital statistics—2003. *Pediatrics*, 115(3):619, 2005.

[77] Klim McPherson, John E Wennberg, Ole B Hovind, Peter Clifford, et al. Small-area variations in the use of common surgical procedures: An international comparison of New England, England, and Norway. *The New England journal of medicine*, 307(21):1310, 1982.

[78] Brian S Mittman. Creating the evidence base for quality improvement collaboratives. *Annals of internal medicine*, 140(11):897–901, 2004.

[79] Naoki Miyagawa, Hiroshi Teramoto, Chun-Biu Li, and Tamiki Komatsuzaki. Decomposability of multivariate interactions. *Complex Systems*, 20(2):165, 2011.

[80] Douglas C Montgomery, Douglas C Montgomery, and Douglas C Montgomery. *Design and analysis of experiments*, volume 7. Wiley New York, 1984.

[81] Jason H Moore. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human heredity*, 56(1-3):73–82, 2003.

[82] Alberto Moraglio and Julian Togelius. Geometric differential evolution. In *Proceedings of the 11th Annual conference on Genetic and evolutionary computation*, pages 1705–1712. ACM, 2009.

[83] L.S. Morales, D. Staiger, J.D. Horbar, J. Carpenter, M. Kenny, J. Geppert, and J. Rogowski. Mortality among very low birthweight infants in hospitals serving minority populations. *American journal of public health*, 95(12):2206, 2005.

[84] Erum Nadeem, S Serene Olin, Laura Campbell Hill, Kimberly Eaton Hoagwood, and Sarah McCue Horwitz. Understanding the components of quality improvement collaboratives: A systematic literature review. *Milbank Quarterly*, 91(2):354–394, 2013.

[85] P.J. Newton, EJ Halcomb, PM Davidson, and A.R. Denniss. Barriers and facilitators to the implementation of the collaborative method: Reflections from a single site. *Quality and Safety in Health Care*, 16(6):409–414, 2007.

[86] I.S. Oh, J.S. Lee, and B.R. Moon. Hybrid genetic algorithms for feature selection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(11):1424–1437, 2004.

[87] World Health Organization et al. Who patient safety curriculum guide for medical schools. 2009.

[88] J. Øvretveit, P. Bate, P. Cleary, S. Cretin, D. Gustafson, K. McInnes, H. McLeod, T. Molfenter, P. Plsek, G. Robert, et al. Quality collaboratives: lessons from research. *Quality and safety in health care*, 11(4):345–351, 2002.

[89] J Øvretveit, Paul Bate, Paul Cleary, Shan Cretin, D Gustafson, K McInnes, H McLeod, T Molfenter, P Plsek, Glenn Robert, et al. Quality collaboratives: lessons from research. *Quality and safety in health care*, 11(4):345–351, 2002.

[90] Ray Pawson and Nick Tilley. *Realistic evaluation*. Sage, 1997.

[91] N.R. Payne, M.J. Finkelstein, M. Liu, J.W. Kaempf, P.J. Sharek, and S. Olsen. Nicu practices and outcomes associated with 9 years of quality improvement collaboratives. *Pediatrics*, 125(3):437–446, 2010.

[92] Martin Pelikan. Analysis of estimation of distribution algorithms and genetic algorithms on NK landscapes. In *Proceedings of the 10th annual conference on Genetic and evolutionary computation*, pages 1033–1040. ACM, 2008.

[93] F. Pernkopf and P. O'Leary. Feature selection for classification using genetic algorithms with a novel encoding. In *Computer Analysis of Images and Patterns*, pages 161–168. Springer, 2001.

[94] Charles Perrow. *Normal Accidents: Living with High Risk Technologies (Updated)*. Princeton University Press, 2011.

[95] Paul E Plsek. Collaborating across organizational boundaries to improve the quality of care. *American journal of infection control*, 25(2):85–95, 1997.

[96] Paul E Plsek and Trisha Greenhalgh. The challenge of complexity in health care. *Bmj*, 323(7313):625–628, 2001.

[97] Paul E Plsek and Tim Wilson. Complexity, leadership, and management in healthcare organisations. *Bmj*, 323(7315):746–749, 2001.

[98] P.E. Plsek. Collaborating across organizational boundaries to improve the quality of care. *American journal of infection control*, 25(2):85–95, 1997.

[99] B.C. Poulton and M.A. West. Effective multidisciplinary teamwork in primary health care. *Journal of Advanced Nursing*, 18(6):918–925, 2008.

[100] M.L. Raymer, W.F. Punch, E.D. Goodman, L.A. Kuhn, and A.K. Jain. Dimensionality reduction using genetic algorithms. *Evolutionary Computation, IEEE Transactions on*, 4(2):164–171, 2000.

[101] Colin Reeves and Christine Wright. An experimental design perspective on genetic algorithms. In *Foundations of Genetic Algorithms 3*, 1995.

[102] Colin R Reeves. Experiments with tuneable fitness landscapes. In *Parallel Problem Solving from Nature PPSN VI*, pages 139–148. Springer, 2000.

[103] Colin R Reeves and Christine C Wright. Epistasis in genetic algorithms: An experimental design perspective. In *Proceedings of the 6th International Conference on Genetic Algorithms*, pages 217–224. Morgan Kaufmann Publishers Inc., 1995.

[104] Ian Reid. Complexity science: Let them eat complexity: the emperor's new toolkit. *BMJ: British Medical Journal*, 324(7330):171, 2002.

[105] D.N. Reshef, Y.A. Reshef, H.K. Finucane, S.R. Grossman, G. McVean, P.J. Turnbaugh, E.S. Lander, M. Mitzenmacher, and P.C. Sabeti. Detecting novel associations in large data sets. *science*, 334(6062):1518–1524, 2011.

[106] A.D. Rodrigues. *Drug-drug interactions*. Informa Healthcare, New York, NY, 2008.

[107] J.A. Rogowski, J.D. Horbar, P.E. Plsek, L.S. Baker, J. Deterding, W.H. Edwards, J. Hocker, A.D. Kantak, P. Lewallen, W. Lewis, et al. Economic implications of neonatal intensive care unit collaborative quality improvement. *Pediatrics*, 107(1):23, 2001.

[108] J.A. Rogowski, J.D. Horbar, D.O. Staiger, M. Kenny, J. Carpenter, and J. Geppert. Indirect vs direct hospital quality indicators for very low-birth-weight infants. *JAMA: the journal of the American Medical Association*, 291(2):202, 2004.

[109] J.A. Rogowski, D.O. Staiger, and J.D. Horbar. Variations in the quality of care for very-low-birthweight infants: implications for policy. *Health Affairs*, 23(5):88–97, 2004.

[110] Jani Rönkkönen, Xiaodong Li, Ville Kyrki, and Jouni Lampinen. A framework for generating tunable test functions for multimodal optimization. *Soft Computing*, 15(9):1689–1706, 2011.

[111] Peter M Rothwell. External validity of randomised controlled trials:to whom do the results of this trial apply?. *The Lancet*, 365(9453):82–93, 2005.

[112] William Rowe, Mark Platt, David C Wedge, Philip J Day, Douglas B Kell, and Joshua Knowles. Analysis of a complete dna–protein affinity landscape. *Journal of The Royal Society Interface*, 7(44):397–408, 2010.

[113] Bill Runciman and Merrilyn Walton. *Safety and ethics in healthcare: a guide to getting it right*. Ashgate Publishing, Ltd., 2007.

[114] E. Salas, N.J. Cooke, and M.A. Rosen. On teams, teamwork, and team performance: Discoveries and developments. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 50(3):540–547, 2008.

[115] T. Sandmann and M. Boutros. Screens, maps & networks: From genome sequences to personalized medicine. *Current Opinion in Genetics & Development*, 22:36–44, 2012.

[116] Elad Schneidman, Susanne Still, Michael J Berry, William Bialek, et al. Network information and connected correlations. *Physical review letters*, 91(23):238701, 2003.

[117] Matthias Schonlau, Rita Mangione-Smith, Kitty S Chan, Joan Keesey, Mayde Rosen, Thomas A Louis, Shin-Yi Wu, and Emmett Keeler. Evaluation of a quality improvement collaborative in asthma care: does it improve processes and outcomes of care? *The Annals of Family Medicine*, 3(3):200–208, 2005.

[118] L.M.T. Schouten, R.P.T.M. Grol, and M.E.J.L. Hulscher. Factors influencing success in quality-improvement collaboratives: Development and psychometric testing of an instrument. *Implementation Science*, 5(1):1–9, 2010.

[119] L.M.T. Schouten, M.E.J.L. Hulscher, J.J.E. Everdingen, R. Huijsman, and R.P.T.M. Grol. Evidence for the impact of quality improvement collaboratives: Systematic review. *Bmj*, 336(7659):1491–1494, 2008.

[120] Loes MT Schouten, MEJL Hulscher, Jannes JE Van Everdingen, Robbert Huijsman, Louis W Niessen, and RPTM Grol. Short-and long-term effects of a quality improvement collaborative on diabetes management. *Implement Sci*, 5:94, 2010.

[121] Paul G Shekelle, Peter J Pronovost, Robert M Wachter, Stephanie L Taylor, Sydney M Dy, Robbie Foy, Susanne Hempel, Kathryn M McDonald, John Ovretveit, Lisa V Rubenstein, et al. Advancing the science of patient safety. *Annals of Internal Medicine*, 154(10):693–696, 2011.

[122] Stephen M Shortell, Jill A Marsteller, Michael Lin, Marjorie L Pearson, Shin-Yi Wu, Peter Mendel, Shan Cretin, and Mayde Rosen. The role of perceived team effectiveness in improving chronic illness care. *Medical care*, 42(11):1040–1048, 2004.

[123] Leif I Solberg. If youve seen one quality improvement collaborative. *The Annals of Family Medicine*, 3(3):198–199, 2005.

[124] Kenneth Tan, Gordon Baxter, Simon Newell, Steve Smye, Peter Dear, Keith Brownlee, and Jonathan Darling. Knowledge elicitation for validation of a neonatal ventilation expert system utilising modified delphi and focus group techniques. *International journal of human-computer studies*, 68(6):344–354, 2010.

[125] Reiko Tanese. *Distributed genetic algorithms for Function Optimization*. PhD thesis, The University of Michigan, Ann Arbor, MI, 1989.

[126] S.L. Taylor, S. Dy, R. Foy, et al. What context features might be important determinants of the effectiveness of patient safety practice interventions? *BMJ Quality & Safety*, 20(7):611–617, 2011.

[127] Dirk Thierens. The linkage tree genetic algorithm. In *Parallel Problem Solving from Nature, PPSN XI*, pages 264–273. Springer, 2010.

[128] Dirk Thierens and Peter AN Bosman. Hierarchical problem solving with the linkage tree genetic algorithm. In *Proceeding of the fifteenth annual conference on Genetic and evolutionary computation conference*, pages 877–884. ACM, 2013.

[129] Nicholas Tomko, Inman Harvey, and Andrew Philippides. Unconstrain the population: The benefits of horizontal gene transfer in genetic algorithms. In *SmartData*, pages 117–127. Springer, 2013.

[130] Shaun Treweek and Merrick Zwarenstein. Making trials matter: pragmatic and explanatory trials and the problem of applicability. *Trials*, 10(37):9, 2009.

[131] M.E. Turner. *Groups at work: Theory and research*. Lawrence Erlbaum, Hillsdale, NJ, 2000.

[132] Ryan J Urbanowicz and Jason H Moore. The application of michigan-style learning classifiersystems to address genetic heterogeneity and epistasisin association studies. In *Proceedings of the 12th annual conference on Genetic and evolutionary computation*, pages 195–202. ACM, 2010.

[133] G. Vaidyanathan. Redefining clinical trials: The age of personalized medicine. *Cell*, 148(6):1079–1080, 2012.

[134] Vesselin K Vassilev, Terence C Fogarty, and Julian F Miller. Information characteristics and the structure of landscapes. *Evolutionary Computation*, 8(1):31–60, 2000.

[135] Edward Weinberger. Correlated and uncorrelated fitness landscapes and how to tell the difference. *Biological cybernetics*, 63(5):325–336, 1990.

[136] John Wennberg and Alan Gittelsohn. Small area variations in health care delivery: A population-based health information system can guide planning and regulatory decision-making. *Science*, 182(4117):1102–1108, 1973.

[137] John E Wennberg. *Tracking Medicine: A Researcher's Quest to Understand Health Care*. Oxford University Press, USA, 2010.

[138] M.A. West. *Effective teamwork: Practical lessons from organizational research*. Blackwell Publishing, Oxford, 2012.

[139] Tim Wilson, Donald M Berwick, and Paul D Cleary. What do collaborative improvement projects do? experience from seven countries. *Joint Commission Journal on Quality and Patient Safety*, 29(2):85–93, 2003.

[140] Tim Wilson, Tim Holt, and Trisha Greenhalgh. Complexity and clinical care. *Bmj*, 323(7314):685–688, 2001.

[141] David D Woods, Leila J Johannesen, Richard I Cook, and Nadine B Sarter. Behind human error: Cognitive systems, computers and hindsight. Technical report, DTIC Document, 1994.

[142] David D Woods, Emily S Patterson, and Richard I Cook. Behind human error: taming complexity to improve patient safety. *Handbook of Human Factors and Ergonomics in Health Care and Patient Safety. London: Lawrence Erlbaum*, pages 459–76, 2007.

[143] Michael Wooldridge. *An introduction to multiagent systems*. John Wiley & Sons, 2009.

[144] Alden H Wright, Richard K Thompson, and Jian Zhang. The computational complexity of NK fitness functions. *IEEE Transactions on Evolutionary Computation*, 4(4):373–379, 2000.

[145] S. Wright. The roles of mutation, inbreeding, crossbreeding and selection in evolution. In *Proceedings of the sixth international congress on genetics*, volume 1, pages 356–366, 1932.

[146] Paul C Young, Gordon B Glade, Gregory J Stoddard, and Chuck Norlin. Evaluation of a learning collaborative to improve the delivery of preventive services by pediatric practices. *Pediatrics*, 117(5):1469–1476, 2006.

[147] Z.J. Yu, F. Haghighat, B. Fung, and L. Zhou. A novel methodology for knowledge discovery through mining associations between building operational data. *Energy and Buildings*, 2011.

[148] J.A.F. Zupancic, D.K. Richardson, J.D. Horbar, J.H. Carpenter, S.K. Lee, G.J. Escobar, et al. Revalidation of the score for neonatal acute physiology in the vermont oxford network. *Pediatrics*, 119(1):e156–e163, 2007.