Team Learning for Healthcare Quality Improvement

NARINE MANUKYAN¹, MARGARET J. EPPSTEIN¹, AND JEFFREY D. HORBAR²

¹Department of Computer Science, University of Vermont, Burlington, VT 05401, USA
²Department of Pediatrics, University of Vermont, Burlington, VT 05405, USA, and Vermont Oxford Network, Burlington, VT 05401, USA

Corresponding author: N. Manukyan (narine.manukyan@uvm.edu)

This work was supported by the NIH Eunice Kennedy Shriver National Institute of Child Health and Human Development under Award 1R21HD068296.

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 because of nonlinear interactions among various demographics, treatments, and practices. In previous studies of collaborations where the goal is a 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 paper, 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 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.

INDEX TERMS Agent-based modeling, collaborative learning, healthcare quality, knowledge diffusion, quality improvement, team collaboration, team learning.

I. INTRODUCTION

Much research has focused on studying team collaboration for joint problem solving [1]–[8]. In this context, diverse teams have been shown to offer some advantage. For example, in [9] 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 [10] 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 [11]–[13]. 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) [14], [15] 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 [16], [17]. 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]–[22]. While there is positive but limited evidence that QICs can result in improved quality of care [23], [24], it is not clear which factors contribute to the effectiveness of teamwork in QICs [25]–[27]. 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 [28], 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 [28], 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 II, we describe the methods used in the ABM portion of the study. In Section III, we show the results of the ABM portion of the study. Section IV discusses data curation and analysis of real hospital data. In Section V, we show the results of the real hospital data analysis. Finally, in Sections VI and Section VII we provide discussion and conclusions.

II. METHODS

A. MODELING THE PROBLEM

We use the same clinical fitness landscape model as used in [28], 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:

$$Pr(s_x) = \left(1 + \exp\left(-\left(\beta_0 + \sum_{i=1}^{N} \beta_i x_i + \sum_{i=1}^{(N-1)} \sum_{j=i+1}^{N} \gamma_{ij} x_i x_j + H\right)\right)\right)^{-1}$$  \hspace{1cm} (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 $\beta_i$ and $\gamma_{ij}$ are randomly drawn from a normal distribution with a mean of 0 and standard deviation of $L^{-0.5}$, where $L$ is the total number of non-zero terms in the model. As in [28] we restrict our landscapes to those with an average fitness of 0.5 ($\beta_0 = 0$), include non-zero coefficients ($\beta_i$) for all main effects, and only model up to two-feature interactions ($\gamma_{ij}$); 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.

B. POPULATION INITIALIZATION

In [28] 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 [28] 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. 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 [28]. 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 [28], we compare search results between populations with initially clustered and scattered distributions. (Note: in section IV we show additional analysis of a larger data set of 20 real-valued clinical practices that further supports the notion of clustered hospitals.)

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. 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 2](attachment:figure2.png)

**Figure 2.** 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. 1b) that is close to that of the real hospital data (Fig. 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. 1c) is very different from that of the real hospitals.

**C. 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 [28]) 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 similarity, so the resulting teams are not necessarily optimally homophilous.)

A visual illustration of the PickSimilarTeams algorithm is shown in Fig. 2b, where the algorithm has divided the population of $M = 50$ individuals shown in Fig. 2a into $T = 5$ teams, with $M_T = 10$ individuals in each team.

### D. TEAM LEARNING

We use the team learning algorithm described in [28], 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 [28], 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 [28] 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 [29]), while also providing agent-specific customized recommendations for change based on where each agent’s fitness lies relative to the others in its team.

### E. 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, the length of the $tabu$ period, as shown in Table 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 1).

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 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 pairwise nHDs within each team. The abbreviations shown in Table 2 are used to label subsequent plots.

### TABLE 1. Factors varied in the ABM simulations.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Values Studied</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Population Type</td>
<td>(Clustered, Scattered)</td>
</tr>
<tr>
<td>Team Formation Mechanism</td>
<td>(Homophilous, Random)</td>
</tr>
<tr>
<td>Team Size ($M_T$)</td>
<td>{1, 2, 4, 5, 10, 20, 25, 50, 100}</td>
</tr>
<tr>
<td>Steps between Team Reformation</td>
<td>{1, 4, 5, 10, 20, 50, 100}</td>
</tr>
<tr>
<td>Noise in Fitness Evaluation</td>
<td>(None, Low, High)</td>
</tr>
<tr>
<td>(corresponding # patients per trial)</td>
<td>{∞, 3(20), 4(0)}</td>
</tr>
<tr>
<td>two-feature interactions</td>
<td>{0, 495, 2475}</td>
</tr>
<tr>
<td>(% possible two-feature interactions)</td>
<td>{0(%), 10(%), 50(%)}</td>
</tr>
<tr>
<td>$tabu$</td>
<td>{2, 5, 10}</td>
</tr>
</tbody>
</table>

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.

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.

### F. 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 assess for statistically significant differences.

### III. RESULTS

Team search consistently significantly outperformed random search ($p < 0.01$, Fig. 3), consistent with [28], although in that study only randomly formed teams were studied.
FIGURE 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.

When averaged over 100 random landscapes, the performance of individual random 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, with 40 patients per trial). Note that the order of performance from highest to lowest fitness shown in Fig. 3 matches with the order of initial within-team similarity shown in Table 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. 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. 4). Homophilous teams of clustered agents (Fig. 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. 4a,c with no noise to 4b,d with high noise) and landscape complexity (compare Fig. 4a,b with no two-feature interactions to 4c,d with 2475 two-feature interactions). One apparent anomaly in these results occurs on complex landscapes with no noise (Fig. 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 congregate 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 reformation increases above 20 (Fig. 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. 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. 5b, black line at 50 trial steps) with a consequent jump in the rate of fitness improvement (Fig. 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. 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. 5b, red line at 50 trial steps) and the rate of learning decreases even more (Fig. 5a, red line after reformation at trial step 50).

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 [28] 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. 6) and within-population similarity (Fig. 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.

When there is no noise, each of the ten teams converged to a single vector (Fig. 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. 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. 7) because as learning progresses the most similar teams tend to stay together (Fig. 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. 6b, upper lines). However, at high noise levels the within-population nHD steadily increases (Fig. 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. 6c), even as fitnesses increase through the learning process (recall Fig. 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.

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. 8). Not surprisingly, learning became easier as noise decreased. What we found more interesting is that the advantage of homophilous teams over random teams was increasingly pronounced with higher noise (fewer patients per trial), especially when starting from clustered initial populations (Fig. 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 [28], 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 trial steps each hospital was forced to wait before reevaluating the same feature on landscapes with intermediate complexity (495 two-feature interactions) (Fig. 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. 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. 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. 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 taboo is low and fitness is noise-free. When teams are never or rarely reformed and taboo 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. 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. 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. 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. 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
FIGURE 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: noise-free or high noise (40 patients per trial), respectively.

FIGURE 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.

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. 1a provides 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 further 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.

IV. 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 listserves, 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) 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.** 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).

<table>
<thead>
<tr>
<th>#</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Prenatal Care</td>
</tr>
<tr>
<td>2</td>
<td>Antenatal Steroids</td>
</tr>
<tr>
<td>3</td>
<td>Vaginal Delivery</td>
</tr>
<tr>
<td>4</td>
<td>Oxygen during initial resuscitation</td>
</tr>
<tr>
<td>5</td>
<td>Face mask ventilation during initial resuscitation</td>
</tr>
<tr>
<td>6</td>
<td>Endotracheal tube during initial resuscitation</td>
</tr>
<tr>
<td>7</td>
<td>Epinephrine during initial resuscitation</td>
</tr>
<tr>
<td>8</td>
<td>Cardiac compressions during initial resuscitation</td>
</tr>
<tr>
<td>9</td>
<td>Cranial imaging-on-before day 23</td>
</tr>
<tr>
<td>10</td>
<td>Oxygen After Initial resuscitation</td>
</tr>
<tr>
<td>11</td>
<td>Conventional ventilation after initial resuscitation</td>
</tr>
<tr>
<td>12</td>
<td>High frequency ventilation after initial resuscitation</td>
</tr>
<tr>
<td>13</td>
<td>Nasal CPAP after initial resuscitation</td>
</tr>
<tr>
<td>14</td>
<td>Surfactant at any time</td>
</tr>
<tr>
<td>15</td>
<td>Steroids for Chronic Lung Disease</td>
</tr>
<tr>
<td>16</td>
<td>Indomethacin for any reason</td>
</tr>
<tr>
<td>17</td>
<td>Retinopathy of Prematurity Surgery</td>
</tr>
<tr>
<td>18</td>
<td>Necrotizing Enterocolitis Surgery</td>
</tr>
<tr>
<td>19</td>
<td>Other Minor Surgery</td>
</tr>
<tr>
<td>20</td>
<td>Any Major Surgery</td>
</tr>
</tbody>
</table>

The number of hospitals in the VON grew from about 50 hospitals in 1990 to more than 800 hospitals in 2010 (see Fig. 11a). Starting in 1995, a small subset of these participated in NICQs (Fig. 11a, black). Of the remainder, some had missing data (Fig. 11a, dark red) and were thus excluded from this study. A more detailed histogram of hospitals that participated in NICQs is shown in Fig. 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. 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 [23], [24]. 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, 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 sub-population) 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).

V. 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. 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.

With the exception of the 1999 NICQ, the within-team nEDs of NICQ hospitals (Fig. 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. 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. 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. 12). The fact that the network is growing over these years (Fig. 12a) undoubtedly contributes to these increases in diversity, so it is
Homophilous teams (picked by within team EDs for either randomly formed teams, real NICQ teams or non-NICQ subpopulations, where nEDs are calculated for either 20 Distances (nEDs) (i.e., subpopulation closeness) in the NICQ and FIGURE 12.

(a) Mean pairwise within subpopulation normalized Euclidean knowledge diffusion (rather than knowledge creation). Environments, like healthcare QICs, where the emphasis is on consistent nature of this finding suggests that homophilous teams may be beneficial in real world collaborative learning.

(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.

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.

VI. 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 learning in teams through collaborative 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. 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 ([30], [31], [32]), we previously [28] 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 fitness 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 II-D. 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 [33]–[35]. 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.

VII. CONCLUSION

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.

ACKNOWLEDGMENT
The authors would like to thank M. Kenny, K. Leahy, and the Vermont Oxford Network for their time and effort in providing the hospital data, and S. Mukherjee for her help in curating the data. We also thank S. A. Kauffman, J. S. Buzas, and D. M. Rizzo for stimulating discussions that benefited this work.

REFERENCES

Narine Manukyan received the B.S. degree in applied mathematics and informatics from Yerevan State University, Yerevan, Armenia, in 2008, and the M.S. degree in computer science from the University of Vermont, Burlington, VT, USA, in 2011, where she is currently pursuing the Ph.D. degree in computer science and serves as a Chapter President of Upsilon Pi Epsilon and Vice President of Education in the Project Management Institute Champlain Chapter. Her research interests include data science, artificial intelligence, evolutionary computation, complex networks, data mining, machine learning, and modeling complex systems.
MARGARET J. EPPSTEIN is an Associate Professor and the Chair of the Department of Computer Science, University of Vermont, Burlington, VT, USA. She received the B.S. degree in zoology from Michigan State University, East Lansing, MI, USA, in 1978, and the M.S. degree in computer science and the Ph.D. degree in environmental engineering from the University of Vermont in 1983 and 1997, respectively. She has been on the Computer Science Faculty with the University of Vermont since 1983 (Lecturer from 1983 to 2001, Research Assistant Professor from 1997 to 2002, Assistant Professor from 2002 to 2008, Associate Professor since 2008) and was the Founding Director of the Vermont Complex Systems Center from 2006 to 2010. Her current research interests involve complex systems analysis and modeling in a wide variety of application domains, including biological, environmental, technological, and social systems.

JEFFREY D. HORBAR is a Board-Certified Neonatologist and Clinical Scientist with extensive experience in clinical research and its application to the improvement of neonatal care. He is currently the Jerold F. Lucey Professor of neonatal medicine with the University of Vermont College of Medicine, Burlington, VT, USA, a Chief Executive and Scientific Officer of the Vermont Oxford Network, an Associate Editor of Pediatrics, and Co-Editor of the Neonatal Review Group of the Cochrane Collaboration. He leads the Vermont Oxford Network Database which is used by over 950 NICUs to monitor and improve outcomes for high risk newborn infants. He is the Director of the Vermont Oxford Network NICQ and iNICQ Improvement Collaboratives in which multidisciplinary teams of health professionals from around the world work together under the guidance of expert faculty to improve the quality and safety of medical care for newborn infants and their families.

* * *