Fuzzy Performance Profiles: Towards Personalized CPR Refresher Training

Florian Sense (florian.sense@infinitetactics.com)
InfiniteTactics, LLC
Dayton, OH, USA

Lauren Sanderson (lauren.sanderson@rqipartners.com)
Joshua Onia (josh.onia@rqipartners.com)
RQI Partners, LLC
Dallas, TX, USA

Michael Krusmark (michael.krusmark.ctr@us.af.mil)
L3Harris Technologies
Melbourne, FL, USA

Joshua Fiechter (josh@kairosresearch.com)
Kairos Research
Dayton, OH, USA

Michael G. Collins (michael.collins.74.ctr@us.af.mil)
ORISE at AFRL
Dayton, OH, USA

Tiffany Jastrzembski (tiffany.myers.1@us.af.mil)
Air Force Research Laboratory, 711th Human Performance Wing
Wright-Patterson Air Force Base, Dayton, OH, USA

Abstract

Cardiopulmonary resuscitation (CPR) is the single greatest determinant in survival from cardiac arrest and is an essential part of every medical professional's toolkit. Effectively responding to cardiac arrests requires CPR proficiency, which is best retained through frequent training. One such model is Resuscitation Quality Improvement (RQI)® that requires users to exhibit proficient performance in compression and ventilation skills every 3 months. While frequent refresher trainings are superior to the traditional 24-month intervals (Cheng et al., 2020), they still need to consider inter-individual differences. Healthcare environments are rife for personalized, adaptive approaches to tracing users' proficiency over time. To this end, we explore how users that perform similarly over time can be clustered together. Such performance profiles could ultimately enhance the benefits of frequent training with personalized efficiency for users with different needs. With the long-term goal of building an adaptive scheduling tool in mind, we present some initial explorations in this domain. Using *k*-means clustering, we show that a small number of clusters seems sufficient to create meaningful performance profiles to make out-of-sample predictions. Furthermore, our simulation study suggests that fuzzy membership to said clusters can be leveraged to enhance predictions. We discuss potential next steps in which these fuzzy performance profiles can be employed by more powerful predictive models to move the field towards personalized, adaptive training schedules that improve learning efficiency with the goal of increasing survival outcomes after cardiac arrest.

Keywords: Cardiopulmonary Resuscitation (CPR); Refresher training; Resuscitation Quality Improvement; Personalized learning; Clustering; Performance predictions

Introduction

The quality of CPR provided by medical professionals has been shown to be inconsistent in both in-hospital (Abella et al., 2005) and out-of-hospital (Wik et al., 2005) cardiac arrests. To improve consistency, various enhancements to how the necessary skills are taught and maintained have been recommended (Cheng et al., 2018). The shift in focus is reflected in the 2020 AHA Guidelines for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care (Cheng et al., 2020). These enhancements to CPR training can improve educational outcomes, patient outcomes, and survival after cardiac arrest (Cheng et al., 2018)—primarily by increasing educational efficiency as a factor in Utstein's formula for survival (Søreide et al., 2013). Increasing educational efficiency includes instructional design features that support a mastery learning model such as the use of feedback (Gruber, Stumpf, Zapletal, Neuhold, & Fischer, 2012; Yeung et al., 2009), spaced learning (Y. Lin, Cheng, Grant, Currie, & Hecker, 2018), assessment, and innovative educational strategies (Cheng et al., 2018). Performance can be scored objectively against clear guidelines (Merchant et al., 2020), which means rich usage logs can be maintained for each user. Thus, CPR training is an ideal domain for technology-enhanced adaptive learning and recent efforts are focused on increased adaptivity (Oermann, Krusmark, Kardong-Edgren, Jastrzembski, & Gluck, 2021).

Here, we will emphasize one dimension of quality CPR: frequency of training. Healthcare providers are trained in

CPR as part of their training and are typically required to retrain on a fixed schedule to maintain compliance. Historically, these trainings have occurred every two years, but recent evidence suggests that more frequent refresher trainings are necessary to prevent skill decay (Cheng et al., 2018; Sutton, Nadkarni, & Abella, 2012). The Resuscitation Quality Improvement (RQI®) program for CPR skills, for example, has participating users complete quarterly skills assessment. The RQI program has been shown to significantly increase users' CPR skills (Kardong-Edgren et al., 2020) but we believe further improvements can be made by personalizing the scheduling of refresher trainings to the needs of individual users. Recent work suggests that adaptively scheduling training intervals can lead to faster acquisition, better retention, and lower costs and risks (Oermann, Krusmark, Kardong-Edgren, Jastrzembski, & Gluck, 2020; Kerfoot, 2010; Oermann et al., 2021).

To aid and inform the transition from fixed to adaptively scheduled retraining intervals, the current study will explore three aspects related to extracting and leveraging performance profiles across training sessions. If similar performance in the past is likely to result in similar performance in the future, predictions can be informed by clusters of similar performance profiles (e.g., because two users make similar mistakes or because knowledge about two skills decays at similar rates). The long-term goal is to leverage meaningful profiles to create personalized refresher training schedules that can replace fixed-interval assessments. To this end, we will explore three research questions pertaining to clustering performances to extract user profiles:

RQ1: How to determine the number of clusters/profiles? The "optimal" number of clusters depends on the purpose of the analysis and is rarely easily quantifiable (Kodinariya & Makwana, 2013). We consider our main goal *predicting* rather than *explaining* (Shmueli, 2010). Therefore, we devised a simulation study on usage logs that focuses on quantifying the out-of-sample prediction error (Yarkoni & Westfall, 2017) to approximate a useful number of performance profiles.

RQ2: Can fuzzy instead of crisp cluster membership improve predictions? We used k-means clustering (Hartigan & Wong, 1979) for our explorations. By default, clustering algorithms assign each input to one of the k clusters. Such *crisp* assignment can be contrasted with *fuzzy* assignment (Ruspini, 1969; Dunn, 1973; Bezdek, 1981) in which each input belongs to all k clusters simultaneously and has membership coefficients γ_k that indicate its relative proximity to each cluster k (see Methods for details and footnote 1 for a numerical example). Crisp assignment effectively reduces all members of a cluster to be identical, while the fuzzy assignment imposes less of a dimensionality reduction. Fuzzy assignment could aid predictions (by pooling important variance and in-between states) or harm predictions (by amplifying noise).

RQ3: Which aspects of performance should be clustered? To maintain compliance in the RQI program, each user needs to be proficient in four related but independent CPR skills: compressions and ventilations performed on an adult and an infant manikin (Merchant et al., 2020). Correlations between skills are positive but not strong enough ($r < 0.4$; see Discussion) to consider performance across skills interchangeable. Consequently, it might make sense to assign each user's performance on a given skill to a distinct cluster (cf. Fig. 1B). However, if scheduling of future refresher trainings was a function of cluster membership, it would make more sense to assign each user's performance across *all* skills to a distinct cluster (cf. Fig. 1A); otherwise, different skills would be scheduled to be retrained at different intervals—that might be most effective but the administrative burden on hospitals might be untenable. To illuminate the potential (dis)advantages of either approach, we will explore and compare both.

In the following, we will describe the dataset we used to explore the above questions. Next, the simulation study's approach will be outlined in detail before the results will be presented. These results and their implications will be discussed in the final section.

Methods

Dataset

The dataset extracted from users participating in the Resuscitation Quality Improvement (RQI®) program for CPR skills in three hospitals in the USA. Users participated every 3 months (quarterly) and we selected a period of 5 quarters (i.e., Q1, Q2, ...).

Users experienced a skills assessment with feedback in all quarters. These measured a user's CPR skills on both an infant and adult manikin on the RQI Simulation Station. In each assessment, users received both auditory and visual real-time feedback on specific metrics of each skill that accumulated to an overall score. Each assessment was followed by a comprehensive debriefing. In Q1 and Q5, a baseline assessment without feedback was required before the skills assessment to measure a user's retention of skill over time. Users were required to pass each skill assessment (i.e., score $\geq 75\%$) to maintain compliance (unlimited number of attempts). The baseline assessment without feedback in Q1 and Q5 did not require passing, but users were provided an overall score and debriefing upon completion.

We only used overall scores on the first attempt in each quarter. For Q1 and Q5, these were baseline assessment scores; for Q2–4, these were skills assessment scores (i.e., without and with feedback respectively). Baseline scores were included as a valid measurement of a user's ability to perform each skill without the aid of real-time feedback. Only the first attempt was included as it is the most informative data point: With passing scores required in every quarter, the last data point for any skill is always in the range 75–100%. Most users passed after the first or second attempt. By including additional attempts, the most discriminating data points were retained while making it easier to apply the out-of-the-box

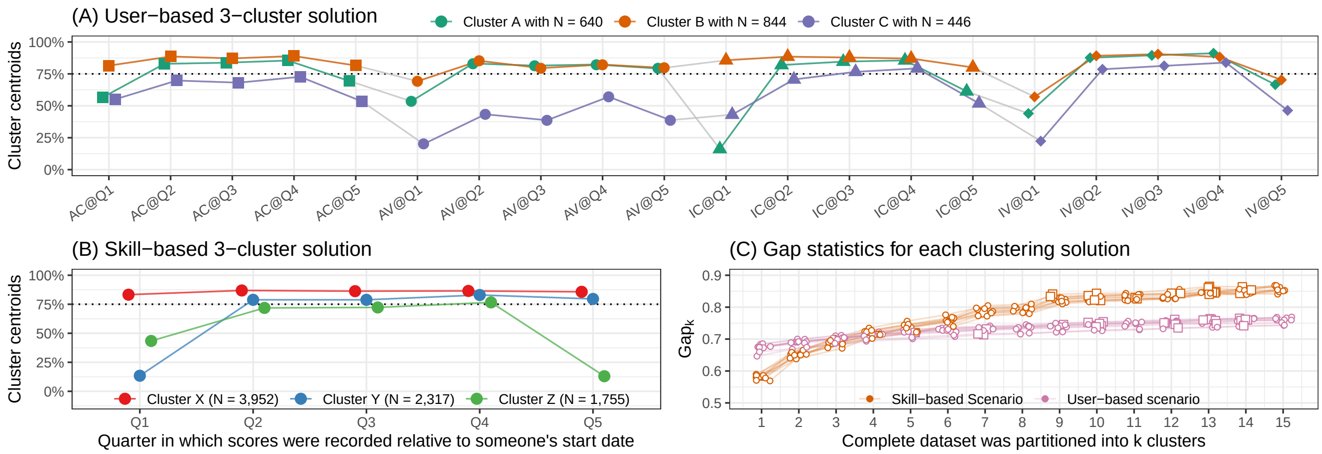

Figure 1: Overview of clustering solutions. **(A)** The user-based approach considers all 20 observations of a user simultaneously when clustering: the four skills (adult compressions, adult ventilations, infant compressions, and infant ventilations) are marked with different shapes across the five quarters (Q1–5). Each user is assigned to one cluster (see N's in the legend). The dotted horizontal line at 75% demarcates the proficiency threshold. **(B)** The skill-based approach considers the five observations (across Q1–5) associated with each skill separately. Each skill a user performs is separately assigned to a cluster (thus larger N's; see legend) **(C)** The Gap statistic (Tibshirani et al., 2001) estimated for $1 \leq k \leq 15$ across 10 iterations for the two approaches. Squares mark the best $\hat{k}$ in each iteration.

implementation of the clustering algorithm.

We included all users with at least two baselines (N = 3,119). Of these, 1,930 (62%) had scores on all skills in all five quarters (i.e., complete user trajectories; Fig. 1A) and there were 8,024 (64%) user-skill combinations with scores on all quarters (i.e., complete skill trajectories; Fig. 1B).

Simulation study

To address our research questions, we conducted a retrospective simulation study on the data described above. The simulation followed these steps for $2 \leq k \leq 15$: **(1)** The data were split into training/test sets (70/30) such that each user was either in the training or test set; **(2)** Using complete trajectories in the training data, k-means clustering was performed; **(3)** Each trajectory in the test data was assigned to one of the k clusters by minimizing the squared Euclidean distance between trajectories and cluster centroids. Only quarters 1–4 were used for cluster assignment; **(4)** *Crisp* predictions are equal to the centroids of the cluster a trajectory was assigned to; **(5)** To make *fuzzy* predictions, each trajectory's membership coefficients to each of the k clusters were computed (see details below). These were used as weights when computing a weighted average across the cluster centroids[1].

[1]For example, suppose someone's adult compression scores are 80 in Q1–4. For the three skill-based clusters in Fig. 1B, that trajectory would yield membership coefficients of 0.888, 0.028, and 0.084 for clusters X, Y, and Z, respectively. The cluster centroids for Q5 are 85.7, 79.6, and 13.0 for X, Y, and Z. The unweighted mean across the three centroids is 59.4; the coefficients-weighted mean is 79.4. Thus, this hypothetical person's *fuzzy* prediction for Q5 would be biased towards cluster X's centroid. Their *crisp* prediction for Q5 would simply be cluster X's centroid (85.7) since that is the cluster they would be assigned to.

We used 10-fold cross-validation across two scenarios that differed in what constitutes a *trajectory*. In **the user-based scenario**, all observations from a single user were considered a single trajectory comprising 20 observations (four skills across five quarters; see Fig. 1A). In **the skill-based scenario**, the four skills were treated independently and each trajectory encompassed five observations (one per quarter; see Fig. 1B). Consequently, a user's performance on all skills constitutes their performance profile (i.e., assigned cluster) in the user-based scenario, while each user can potentially have multiple performance profiles in the skill-based scenario.

On every iteration of step (2), k-means clustering (Hartigan & Wong, 1979) as implemented in `R` 3.6.3 (R Core Team, 2020) was performed with 20 random starts. We chose not to impute missing data to avoid amplifying patterns already in the data. At least 60% of trajectories are complete so each training set should contain sufficient observations to partition the data into k clusters.

Computing fuzzy membership coefficients in step (5) comprised two steps: First, each trajectory's distance d from the k centroids c is computed as $d_k = \sum((x - c_k)^2)^{0.5}$, where x is the trajectory. Then, the membership coefficient γ to cluster k is computed by scaling d_k relative to all distances $\mathbf{D}$: $\gamma_k = [\sum(d_k^2/\mathbf{D}^2)^{(1/(m-1))}]^{-1}$ with "fuzzifier" $m = 2$ (Cebeci, 2019). This approach assures that a trajectory's k coefficients sum to 1.

For each iteration, prediction errors were computed separately for (i) quarters 1–4 (used to assign trajectories to clusters; step (3)) and quarter 5 (predictions), and (ii) crisp and fuzzy predictions. Specifically, two types of error were calculated: The root-mean-squared-error (RMSE) and the area under the ROC curve (AUC; (Fawcett, 2006)). The RMSE expresses the prediction error in units of the original scale, which is helpful in gauging how close predicted performance is, on average, to actual performance (lower values are better). The AUC was computed by dichotomizing actual performance as above/below the 75% performance threshold. Hence, the AUC quantifies how well crisp/fuzzy cluster membership can classify whether predicted performance will be proficient or not (higher values are better).

This approach allowed us to address our research questions by (I) checking which value of k minimizes prediction errors (to determine optimal cluster size), (II) comparing crisp and fuzzy predictions on the same data (to ascertain the potential advantage of fuzzy predictions), and (III) directly comparing user- and skill-based clustering solutions.

Results

Figure 2 summarizes the results of the simulation study and pertains to all three research questions outlined in the Introduction. The figure shows the two prediction errors AUC (top row) and RMSE (bottom row) for the skill-based (left column) and user-based (right column) scenario for $2 \leq k \leq 15$. Each panel shows four lines: Dashed lines are the errors computed across quarters 1–4, which were used to make crisp cluster assignments and compute the fuzzy membership coefficients; solid lines connect errors for the fifth quarter, which was withheld completely and thus amounts to a real prediction. The two types of prediction that were made—crisp and fuzzy—are color-coded. With this figure as our guide, we now consider the three research questions in turn.

First, for both skill- and user-based clusters, the optimal number of clusters is small. Looking at the solid lines in Fig. 2, we see that little improvement in achieved beyond $k \approx 3$. The only exception might be making binary predictions in the fuzzy user-based scenario since the AUC increases up to 8 or 10 clusters. The corresponding decrease in RMSE, however, is minimal, which indicates absolute accuracy does not increase in lockstep with classification accuracy. The RMSE in the skill-based scenario (lower left panel) shows the most drastic divergence between the crisp and fuzzy approaches: For crisp predictions in Q5, any $k > 2$ produces markedly larger errors, while the corresponding fuzzy predictions clearly improve if k is increased from 2 to 3 (both in AUC and RMSE).

The 3-cluster solutions for the user- and skill-based scenarios are shown in Fig. 1A and B, respectively. The skill-based clusters in particular capture profiles that clearly make sense: consistently high (X), proficient after Q1 (Y), and only proficient with feedback enabled (Z). Importantly, each user's performance on the four skills could be assigned to different clusters/profiles. For the user-based scenario, on the other hand, the three patterns are more subtle (see Fig. 1A) and simultaneously affect all four skills for a user. However, again, sensible patterns are apparent: Users only proficient with feedback enabled (A), consistently proficient (B), and inconsistent/struggling throughout (C). We also see an overall trend across the user-based clusters that highlights decreased performance without live feedback (Q1 and Q5).

Second, there is a consistent advantage of predictions using fuzzy rather than crisp cluster membership. In Fig. 2, the blue solid lines are always higher than the orange solid lines for the AUC (top row) and lower for the RMSE (bottom row). This difference is particularly stark in the skill-based scenario, where fuzzy predictions are vastly better than their crisp counterparts. An additional advantage might be that fuzziness better protects against over-fitting (analogous to, e.g., Bayesian model averaging (Hoeting, Madigan, Raftery, & Volinsky, 1999) or regularization (Hastie, Tibshirani, & Friedman, 2009)). Here, over-fitting would be apparent if the dashed and solid lines diverge as k increases, which is the case for crisp (orange) but not fuzzy (blue) predictions.

Third, predictions are better if individual skills are clustered rather than a user's entire performance history. Comparing the solid blue lines between the scenarios (left vs. right column in Fig. 2), we see higher AUC and lower RMSE for the skill-based scenario. The reverse is true for orange lines, which reflects the general advantage of fuzzy predictions outlined in the previous paragraph. Hence, these results suggest that using cluster membership alone, the most accurate predictions are made using skill-specific trajectories' fuzzy memberships to 3 clusters.

To complement the results directly relevant to our research questions, we also included a short exploration of what might be considered the optimal number of clusters outside a prediction framework. To this end, we computed the Gap statistic (Tibshirani et al., 2001) as implemented in (Maechler, Rousseeuw, Struyf, Hubert, & Hornik, 2019) for $1 \leq k \leq 15$ using 250 bootstrap samples on ten iterations each with random subsets of 80% of users. Results are summarized in Fig. 1C, showing the estimated Gap_k values (the ten runs are grouped by lines, color-coded for approach). Again, we contrast the user- and skill-based approaches and, following (Tibshirani et al., 2001), define the best solution as the smallest k for which $Gap(k) \geq Gap(k+1) - s_{k+1}$, where s is the standard error. For each iteration in Fig. 1C, those instances are highlighted as squares, which fall between $5 \leq k \leq 14$ for both approaches. Thus, if our goal was to explain rather than predict (Shmueli, 2010), a larger number of clusters might be appropriate. However, our simulations showed that with prediction as our main goal, a small number is sufficient and mistakenly using the results of this explanatory analysis for prediction purposes would result in over-fitting.

Discussion

Our simulation on usage logs of medical professionals' CPR refresher training suggests that a relatively small number of performance profiles is sufficient to make decent predictions. These predictions were consistently more accurate if membership to each performance profile was fuzzy rather than crisp and if skills were considered independently.

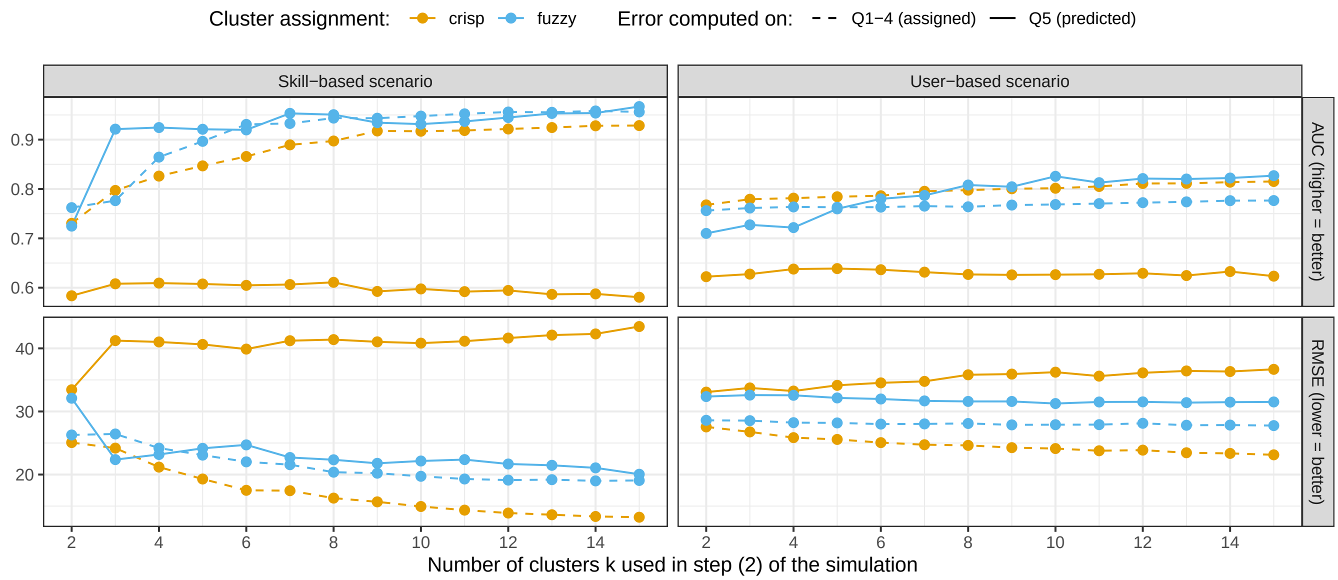

Figure 2: Prediction errors AUC and RMSE in the test sets. Comparing errors computed on Q1–4 (assigned) and Q5 (predicted) for the crisp (orange) and fuzzy (blue) predictions. Detailed description in the first paragraph of the Results section.

Contextualizing these findings, it should be acknowledged that this a difficult prediction task. The only information used in our current explorations are the overall scores of first attempts of each learner's mandatory, quarterly skills assessments. Nevertheless, membership to a small number of clusters affords above-chance predictions. Predictions are considerably better if (a) skills are considered independently, (b) memberships are "fuzzified", and (c) we only want to predict proficiency (i.e., above/below threshold) rather than exact scores (see Fig. 2). One potential explanation is that differences in learning profiles stem from different sources that are each captured by different clusters. Fuzzy memberships allow for those sources to be pooled to make predictions, while crisp profiles more rigidly represent a single source of variance. The promising results in the current exploration suggest a number of extensions that should further improve the predictive accuracy. We will discuss some of these below.

Lower-level scores, exact timestamps, and additional information about users could be leveraged to greatly enrich the data. Each skills' overall score is comprised of several sub-scores (e.g., compression rate, compression depth, hand positioning, etc.). Providing personalized feedback based on sub-scores has proved effective (Yeung et al., 2009; Cheng et al., 2015), suggesting that sub-scores could be leveraged for fine-tuning predictions. Furthermore, demographic information such as body mass index have been linked to quality CPR (Sayee & McCluskey, 2012; C.-C. Lin et al., 2016; López-González et al., 2016) and it is likely differences in work setting within a hospital would be relevant, too (e.g. emergency care vs. outpatient clinic).

To model such greatly enriched data, more sophisticated models would need to be deployed. Machine learning algorithms for supervised learning would be the obvious choice since they have an unparalleled ability to learn arbitrary statistical regularities in sufficiently large, annotated data (Hastie et al., 2009). Such models—especially the "deep" kind (LeCun, Bengio, & Hinton, 2015; Botvinick, Wang, Dabney, Miller, & Kurth-Nelson, 2020)—are great at making performance predictions but it is often impossible to determine how a prediction was arrived at (Gunning, 2017). The cluster-based approach presented here is the polar opposite—for the crisp case, predictions directly follow from someone's proximity to a given cluster; for the fuzzy case, predictions are weighted by someone's relative distance from each cluster. A promising middle ground between these two extremes—models with and without assumptions (Pelánek, 2017, Fig. 4)—are cognitive models. These use theoretical assumptions to constrain mathematical relationships between in- and outputs and rely on a small number of parameters that usually map onto domain-relevant concepts (McClelland, 2009). Promising examples for the current domain are knowledge tracing models (Corbett & Anderson, 1994; Yudelson, Koedinger, & Gordon, 2013) and learning-specific process models (Pavlik & Anderson, 2008; Jastrzembski, Gluck, & Gunzelmann, 2006; Walsh, Gluck, Gunzelmann, Jastrzembski, & Krusmark, 2018; Walsh, Gluck, Gunzelmann, Jastrzembski, Krusmark, Myung, et al., 2018). These cognitive models would trace acquisition and retention of skills over time in a much more nuanced way and could be augmented with cluster membership information (Ayers, Nugent, & Dean, 2008; Nedungadi & Remya, 2014; Sense, Collins, Krusmark, & Jastrzembski, 2020).

Performance on the four skills is positively correlated but correlations are small[2] enough that a single user regularly obtains very different scores. The lack of strong correlations between skills explains why the user-based clusters (Fig. 1A) yield less discriminating profiles. With this in mind, using the skill-based clustering approach seems sensible. One implication of these patterns is that the most efficient training regime would consider skills independently, with personalized schedules devised for each skill separately.

Ultimately, the goal is to transition from fixed training intervals to fully personalized, adaptive schedules. The results presented here support these efforts and will contribute to building predictive models. However, retrospective simulation studies are not sufficient evidence—we will only know whether a given system works once it is used in a hospital setting. Conducting such field studies is costly and should be informed by well-founded assumptions. Recent efforts in this domain have shown promising results when using a cognitive model to adaptively schedule training intervals for nursing students acquiring CPR skills (Jastrzembski et al., 2017; Oermann et al., 2020, 2021).

To summarize, fuzzy membership to a small number of clusters proved beneficial in constructing predictive performance profiles for CPR refresher trainings. We expect that the current results will generalize to future scenarios in which we aim to insert specialized cognitive models into the predictive pipeline. The gold standard for assessing the validity of any predictive system would be a field study in which learners pursue personalized, adaptive schedules prescribed by the model—until then, we believe the simulation approach underpinning our current findings provides a good framework for evaluating the applied potential of models and pitching them against each other. It is our hope that such technology-enhanced adaptive scheduling systems can help medical professionals acquire and retain crucial skills that ultimately translate to increased survival outcomes after cardiac arrest.

Acknowledgements

We would like to acknowledge the cooperative research and development partnership between the Air Force Research Laboratory and RQI Partners that produced the data described in this research.

References

Abella, B. S., Alvarado, J. P., Myklebust, H., Edelson, D. P., Barry, A., O'Hearn, N., … Becker, L. B. (2005). Quality of cardiopulmonary resuscitation during in-hospital cardiac arrest. *Jama*, *293*(3), 305–310.

Ayers, E., Nugent, R., & Dean, N. (2008). Skill set profile clustering based on student capability vectors computed from online tutoring data. *Educational Data Mining 2008*.

Bezdek, J. C. (1981). Objective function clustering. In *Pattern recognition with fuzzy objective function algorithms* (pp. 43–93). Springer.

[2] The weakest correlation is between adult ventilations and infant compressions ($r = 0.202$); the strongest correlation is between infant compressions and ventilations ($r = 0.355$).

Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J., & Kurth-Nelson, Z. (2020). Deep reinforcement learning and its neuroscientific implications. *Neuron*.

Cebeci, Z. (2019). Comparison of internal validity indices for fuzzy clustering. *Journal of Agricultural Informatics*(2), 1-14. doi: 10.17700/jai.2019.10.2.537

Cheng, A., Brown, L. L., Duff, J. P., Davidson, J., Overly, F., Tofil, N. M., … others (2015). Improving cardiopulmonary resuscitation with a cpr feedback device and refresher simulations (cpr cares study): a randomized clinical trial. *JAMA pediatrics*, *169*(2), 137–144.

Cheng, A., Magid, D. J., Auerbach, M., Bhanji, F., Bigham, B. L., Blewer, A. L., … others (2020). Part 6: resuscitation education science: 2020 american heart association guidelines for cardiopulmonary resuscitation and emergency cardiovascular care. *Circulation*, *142*(16_Suppl_2), S551–S579.

Cheng, A., Nadkarni, V. M., Mancini, M. B., Hunt, E. A., Sinz, E. H., Merchant, R. M., … others (2018). Resuscitation education science: educational strategies to improve outcomes from cardiac arrest: a scientific statement from the american heart association. *Circulation*, *138*(6), e82–e122.

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, *4*(4), 253–278.

Dunn, J. C. (1973). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, *3*, 32–57.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, *27*(8), 861–874.

Gruber, J., Stumpf, D., Zapletal, B., Neuhold, S., & Fischer, H. (2012). Real-time feedback systems in cpr. *Trends in Anaesthesia and Critical Care*, *2*(6), 287–294.

Gunning, D. (2017). Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, *2*(2).

Hartigan, J. A., & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, *28*(1), 100–108.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science*, 382–401.

Jastrzembski, T. S., Gluck, K. A., & Gunzelmann, G. (2006). *Knowledge tracing and prediction of future trainee performance* (Tech. Rep.). Florida State Univ Tallahassee Dept of Psychology.

Jastrzembski, T. S., Walsh, M., Krusmark, M., Kardong-Edgren, S., Oermann, M., Dufour, K., … others (2017). Personalizing training to acquire and sustain competence through use of a cognitive model. In *International conference on augmented cognition* (pp. 148–161).

Kardong-Edgren, S., Oermann, M. H., Jastrzembski, T. S., Krusmark, M. A., Gluck, K. A., Molloy, M. A., … Sarasnick, J. A. (2020). Baseline cardiopulmonary resuscitation skill performance of nursing students is improved after one resuscitation quality improvement skill refresher. *Journal for Nurses in Professional Development*, *36*(2), 57–62.

Kerfoot, B. P. (2010). Adaptive spaced education improves learning efficiency: a randomized controlled trial. *The Journal of urology*, *183*(2), 678–681.

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of cluster in k-means clustering. *International Journal*, *1*(6), 90–95.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Lin, C.-C., Kuo, C.-W., Ng, C.-J., Li, W.-C., Weng, Y.-M., & Chen, J.-C. (2016). Rescuer factors predict high-quality cpr—a manikin-based study of health care providers. *The American journal of emergency medicine*, *34*(1), 20–24.

Lin, Y., Cheng, A., Grant, V. J., Currie, G. R., & Hecker, K. G. (2018). Improving cpr quality with distributed practice and real-time feedback in pediatric healthcare providers–a randomized controlled trial. *Resuscitation*, *130*, 6–12.

López-González, A., Sánchez-López, M., Garcia-Hermoso, A., López-Tendero, J., Rabanales-Sotos, J., & Martínez-Vizcaíno, V. (2016). Muscular fitness as a mediator of quality cardiopulmonary resuscitation. *The American journal of emergency medicine*, *34*(9), 1845–1849.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2019). cluster: Cluster analysis basics and extensions [Computer software manual]. (R package version 2.1.0 — For new features, see the 'Changelog' file (in the package source))

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, *1*(1), 11–38.

Merchant, R. M., Topjian, A. A., Panchal, A. R., Cheng, A., Aziz, K., Berg, K. M., … of Care Writing Groups, S. (2020). Part 1: Executive summary: 2020 american heart association guidelines for cardiopulmonary resuscitation and emergency cardiovascular care. *Circulation*, *142*(16_Suppl_2), S337–S357.

Nedungadi, P., & Remya, M. S. (2014, October). Predicting students' performance on intelligent tutoring system – Personalized clustered BKT (PC-BKT) model. In *2014 IEEE Frontiers in Education Conference (FIE) Proceedings* (pp. 1–6). Madrid, Spain: IEEE. doi: 10.1109/FIE.2014.7044200

Oermann, M. H., Krusmark, M. A., Kardong-Edgren, S., Jastrzembski, T. S., & Gluck, K. A. (2020). Training interval in cardiopulmonary resuscitation. *Plos one*, *15*(1), e0226786.

Oermann, M. H., Krusmark, M. A., Kardong-Edgren, S., Jastrzembski, T. S., & Gluck, K. A. (2021). Personalized training schedules for retention and sustainment of cpr skills. *Simulation in Healthcare*.

Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, *14*(2), 101.

Pelánek, R. (2017). Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. *User Modeling and User-Adapted Interaction*, *27*(3-5), 313–350.

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Ruspini, E. H. (1969). A new approach to clustering. *Information and control*, *15*(1), 22–32.

Sayee, N., & McCluskey, D. (2012). Factors influencing performance of cardiopulmonary resuscitation (cpr) by foundation year 1 hospital doctors. *The Ulster medical journal*, *81*(1), 14.

Sense, F., Collins, M. G., Krusmark, M., & Jastrzembski, T. (2020). Using k-means clustering for out-of-sample predictions of memory retention. In *Proceedings of the 42nd annual conference of the cognitive science society.* Virtual conference.

Shmueli, G. (2010). To explain or to predict? *Statistical science*, *25*(3), 289–310.

Søreide, E., Morrison, L., Hillman, K., Monsieurs, K., Sunde, K., Zideman, D., … others (2013). The formula for survival in resuscitation. *Resuscitation*, *84*(11), 1487–1493.

Sutton, R. M., Nadkarni, V., & Abella, B. S. (2012). "putting it all together" to improve resuscitation quality. *Emergency Medicine Clinics*, *30*(1), 105–122.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423.

Walsh, M. M., Gluck, K. A., Gunzelmann, G., Jastrzembski, T., & Krusmark, M. (2018). Evaluating the theoretic adequacy and applied potential of computational models of the spacing effect. *Cognitive science*, *42*, 644–691.

Walsh, M. M., Gluck, K. A., Gunzelmann, G., Jastrzembski, T., Krusmark, M., Myung, J. I., … Zhou, R. (2018). Mechanisms underlying the spacing effect in learning: A comparison of three computational models. *Journal of Experimental Psychology: General*, *147*(9), 1325.

Wik, L., Kramer-Johansen, J., Myklebust, H., Sørebø, H., Svensson, L., Fellows, B., & Steen, P. A. (2005). Quality of cardiopulmonary resuscitation during out-of-hospital cardiac arrest. *Jama*, *293*(3), 299–304.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

Yeung, J., Meeks, R., Edelson, D., Gao, F., Soar, J., & Perkins, G. D. (2009). The use of cpr feedback/prompt devices during training and cpr performance: a systematic review. *Resuscitation*, *80*(7), 743–751.

Yudelson, M. V., Koedinger, K. R., & Gordon, G. J. (2013).

Individualized bayesian knowledge tracing models. In *International conference on artificial intelligence in education* (pp. 171–180).