Visualizing Clinical Assessment Variation for Quality Improvement: An Example in Preoperative Physical Status Classification

Andrew Wen MS1, Sungrim Moon PhD1, Jungwei Fan PhD1,2
1Division of Digital Health Sciences, Department of Health Sciences Research, 2Kern Center for the Science of Health Care Delivery, Mayo Clinic, Rochester, Minnesota

Introduction
Clinical practice variation is a common phenomenon where clinicians perform differently in assessment, diagnosis, treatment, or prevention. As a gateway to cascaded decisions and dispositions, clinical assessment variation (CAV) has profound influence on the quality and cost in care delivery. For example, inaccurate assessment can lead to misclassification of the disease type, resulting in suboptimal choice of therapy and resource waste. The reasons for CAV are known to be highly diverse, posing difficulty in resolution. A fundamental step prior to tackling any complex CAV problem is being able to detect it in a sensible way. We believe modern visualization techniques hold terrific potential to addressing this challenge by meaningful presentation of the data. In principle, the pertinent factors and degree of variation both need to be rendered into a cognitive map for intuitive scrutiny. The representation should also suggest which variants are desirable and point out a direction for improvement.

To demonstrate the usefulness of a framework that we propose for visualizing CAV, the American Society of Anesthesiologist Physical Status Classification (ASA scoring) was picked as the example application in this study. ASA scoring is commonly used in the stratification of a patient’s fitness before surgery, by considering a range of preoperative factors such as major comorbidities. A score of 1 indicates the healthiest end and 6 the opposite. Despite its wide adoption, ASA scoring exhibits significant CAV with reported inter-rater agreements as high as a moderate 0.61. Our methods used deep learning to represent the preoperative medical history of each surgical case into a dense vector. The dense vectors were learned by modelling the nonlinear and progressive correlations in the original feature space. Unsupervised clustering was applied to reveal comorbidly similar groups based on the rich representation. By color-coding the ASA ordinal scores on top of the clusters with a 2D plot, we were able to intuitively summarize the spatial distribution of CAV. Moreover, the framework allowed inspection into the cluster contents, which greatly facilitate audits of the scoring consistency and score recommendations through cases with similar medical histories.

Methods
The current analytics pipeline involves four major steps: representation encoding, unsupervised clustering, visualizing CAV, and inspection of the clusters. Implementation in the ASA scoring is elaborated as follows:

Step 1: Deep encoding of longitudinal medical history for each surgical case. The study cohort consisted of 83,908 surgical cases corresponding to 67,020 patients from the Mayo Clinic data warehouse between 12/1/2018 and 7/28/2019 (IRB #18-011585). The sequence of encounters and their ICD-10 CM codes within past 6 months were retrieved for each surgical case. The Med2Vec program was then applied over the cohort to learn dense vectors (dimension=200) for representing a rich semantic space of the progressive comorbidities.

Step 2: Dimension reduction and unsupervised clustering of the surgical cases. The t-SNE algorithm was used to model the nonlinear (dis)similarity of the surgical cases and map them onto a 2D plane. The inputs were the Step 1 dense vectors and a perplexity=90 was empirically tuned. The clusters would reflect the spatial distribution of preoperative comorbidities of the patients in a distilled space, without using the known ASA scores as any part of the input features.

Step 3: Visualizing the ASA scores on top of the comorbidity clusters. We color-coded the past clinician-assigned ASA scores on the surgical cases and rendered a 2D plot combing both aspects of information. The superimposed view was meant to reveal the scattering pattern and association of the scoring with the comorbidity subgroups.

Step 4: Determining cluster boundaries and inspection into the cluster contents. The DBSCAN algorithm was used to assign membership of adjacent cases into a definite cluster. This allowed quantifying cluster purity (consistency in scoring) and identifying the representative diagnoses for each cluster. The purity here is defined as an adjusted distribution across the scores within a cluster, computed by first normalizing the in-cluster score distribution over the whole-cohort score distribution, and then renormalizing to a sum of 1 within the cluster.

Results
The clusters by t-SNE are shown in Figure 1, with the temperature hue corresponding to ASA score severity. We can see that close-color cases did tend to cluster together, suggesting that comorbidities drove the scoring decisions as they were supposed to. On the other hand, there were noisy areas where the scores appear diffuse.

![Figure 1. Preoperative comorbidity-clustered surgical cases and their color-coded ASA scores](image)

From Step 4 we found 146 distinct clusters, with a varying degree of purity. In Table 1, we give an example of a highly consistent and inconsistent cluster respectively. The purer cluster is represented by conditions relating to severe reduction of ejection fraction, consistent with the score=4 criterion in published guidelines. The noisier cluster consists of general chronic conditions that are subject to diffuse interpretation (and thus widespread scoring).

<table>
<thead>
<tr>
<th>Top frequent ICD-10 CM codes</th>
<th>ASA score in-cluster distribution (obtained from Step 4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>142.2, 142.1, 134.0, 151.7, Z45.02</td>
<td>0% 1.04% 13.60% 85.36% 0.00% 0%</td>
</tr>
<tr>
<td>I10, E78.5, E11.9, Z00.00, Z12.11</td>
<td>2.76% 24.88% 31.41% 19.65% 21.31% 0%</td>
</tr>
</tbody>
</table>

**Discussion**

We propose an intuitive approach to visualizing CAV and revealing the source of variation. By superimposing the (assumed) response variable and clusters learned according to the explanatory variables, a color-coded 2D scatter plot is informative for representing the spatial proximity and class composition of the clusters. Without loss of generality, the example on ASA scoring demonstrated the approach can assist in auditing inconsistent human classification and reconciling the scores of adjacent surgical cases that share similar preoperative comorbidities. Inspection into the cluster contents did find that relevant ICD codes tend to be grouped together. The noisy clusters remain challenging by nature, but the clusters with high scoring agreement could be curated to refine practice guidelines. Our ongoing work focuses on the development of an interactive frontend that allows our clinician collaborators to conduct a formal evaluation of the cluster quality and utility for scoring assistance. Integration into EMR will be desirable ultimately.

**References**