Data Challenge for a Digital Health Company – Time Limit: 40 Hours
Submitted by: Kalina Slavkova (kslav@sas.upenn.edu)

Aim 1: Assess relationship between cigarettes smoked per day and the prevalence of diabetes.

Datasets:

1. **patient_clinical.csv**: 4231 observations (cases) of 9 variables.
2. **patient_demographics.csv**: 4240 observations of 4 variables. Gender, age, and education are reported, but the meaning of the education values are not defined, so it would be difficult to interpret any findings related to education without knowing what these education values mean.
3. **patient_history.csv**: 4229 observations of 4 variables. This table reports whether the subject currently smokes, takes BP meds, smokes some number of cigarettes per day, has had a stroke, has hypertension, and has diabetes.
4. **patient_notes.csv**: 2611 observations of 1 variable, namely notes from a clinician during an exam. These notes vary greatly.

Data Preprocessing:

1. Merge the clinical, demographics, and history datasets into one data frame
2. Remove cases with NaN values for any of the variables, yielding a final N=3640 patients.
3. Note: If we need to analyze patient notes, we need to consider that it would severely decrease our N value since patient notes are only available for 2611 patients.

Assumptions:

1. The category *diabetes* includes both type I and type II diabetes. There was no discrimination between these two types in the spreadsheet.
2. *TenYearCHD* refers to prevalence of coronary heart disease at 10 year follow up.
3. We want to use as large of a sample size as possible, only excluding cases if they were missing clinical, demographic, or history information.
4. For the purposes of this challenge, only the relationship between diabetes (*diabetes*) and cigarettes smoked per day (*cigsPerDay*) was evaluated.
5. Because no site information was provided, the assumption is that no harmonization over batch variables was required for the purposes of this challenge.
6. All relevant information on cigarettes per day and diabetes were provided in the appropriate field in patient history. Cigarettes per day is assumed to be a patient reported average of the number of cigarettes they smoke in a typical day.
7. The assumption is also that patients who smoke have been smoking for a substantial amount of time before the study that gathered information on the reported averages of cigarettes per day, leading to the consideration of the hypothesis that smoking is a risk factor for diabetes.

Hypothesis: smoking is a risk factor for diabetes, and diabetes incidence is positively correlated with the number of cigarettes smoked per day.

Data Challenge for a Digital Health Company – Time Limit: 40 Hours
Submitted by: Kalina Slavkova (kslav@sas.upenn.edu)

Observation: Among the 3640 cases that for whom we have clinical, demographic, and historic inform, we observe 553 with CHD at 10 year follow up and only 99 who are diabetic among 2085 who are current smoker/smoke at least 1 cigarette daily. This tells us it might be hard to do classification tasks or design good ML models if we wanted to predict diabetes given smoking status. A good ratio of classes is roughly 1:10-1:20 per predictor, so 553 could be enough for building models for CHD prediction, but 99 diabetic cases (roughly 1:33) will require some bias correction and/or class imbalance remedies for such a prediction task. In any case, *we only want to know if there is a correlation between cigarettes smoked per day and diabetic prevalence*.

Additionally, we only have a binary outcome for diabetes, which does not tell us the following

1. What type of diabetes the patient has
2. When the diabetes onset began
3. How diabetes onset relates to smoking habits. Namely, was the patient smoking for a consistent amount of time before diabetes onset? Is cigarettes smoked per day reported as an average of smoking habits before and after diabetes onset, if such habits existed to begin with? The patient notes do not provide information on longitudinal smoking habits.

Visualization:

First, let's quickly look at the distribution of *cigsPerDay* over total population:

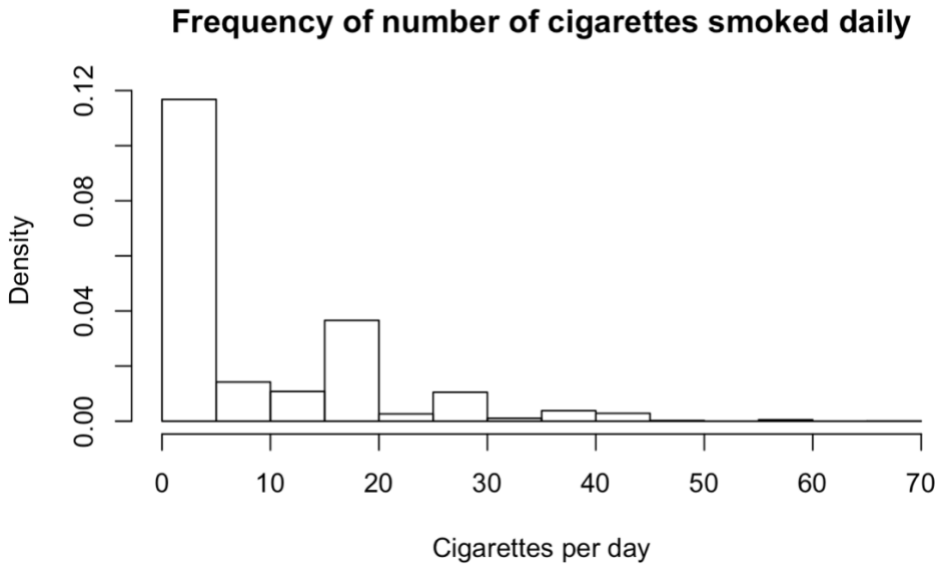


Figure 1. Cigarettes smoked per day over the entire patient population. There are 1780 smokers, and 1860 non-smokers.

Next, let's plot the cigarettes smoked by day when stratified by diabetic status. We will plot a density plot so that the two distributions are on the same scale.

Data Challenge for a Digital Health Company – Time Limit: 40 Hours
Submitted by: Kalina Slavkova (kslav@sas.upenn.edu)

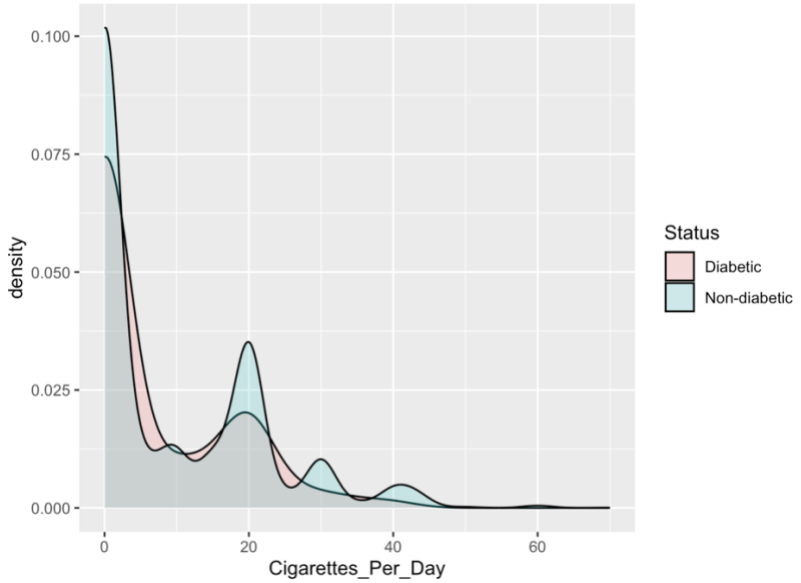


Figure 2. Cigarettes smoked by day by diabetic (red) and non-diabetic (blue) populations. Among diabetics, there are 63 non-smokers and 36 smokers. Among non-diabetics, there are 1744 smokers and 1797 non-smokers.

Observation: Both of the distributions in Figure 2 are skewed (non-normal distributions).

Statistical Analysis:

It's clear from our plot that our two distributions are not normally distributed. We can use a **non-parametric test** to determine whether the distributions of our two vectors of data with different-lengths are significantly different. An appropriate choice is the **Wilcoxon rank sum test** (Mann-Whitney U), which is the non-parametric version of the unpaired t-test.

Null hypothesis: The null hypothesis in our case is that the number of cigarettes smoked per day in the diabetic group is not larger than the number of cigarettes smoked per day in the non-diabetic group. We will use $p<0.05$ as the cut-off level of significance.

Alternative hypothesis 1: Our alternative hypothesis is that the number of cigarettes smoked per day is greater in the diabetic group than the number of cigarettes smoked per day in the non-diabetic group. The assumption is that smoking is harmful and is a risk factor in the development of diabetes, and the more cigarettes a person smokes, the greater the risk and prevalence of diabetes. Thus, we want to evaluate a one-sided *p*-value using the Wilcoxon rank sum test.

- *Result:* ***p*= 0.9926**, we accept the null hypothesis.
- *Discussion*: It seems that the distribution of cigarettes smoked per day in the diabetic group does not significantly skew larger than the cigarettes smoked per day in the non-diabetic group.

Alternative hypothesis 2: What if we test the alternative hypothesis that the diabetic group actually smokes less cigarettes per day than the non-diabetic group? This is based on the possibility that patients with diabetes smoke less due to proper disease management and development of healthier habits, like cutting back on smoking.

- *Result*: ***p*=0.0074**, we reject the null hypothesis and accept the 2nd alternative hypothesis.
- *Discussion*: The distribution of number of cigarettes smoked per day in the diabetic group is significantly less in terms of median compared to the number of cigarettes smoked per day in the non-diabetic group. This could suggest that patients with diabetes smoke less potentially due to proper disease management and development of healthier habits, like cutting back on smoking. More information must be collected, like longitudinal assessment of smoking over time and onset of positive diabetic status.

Aim 2: Design a segmentation algorithm for segmenting retinal vasculature from ophthalmic images, using two sets of ground truth segmentations for evaluation.

Datasets:

1. **Grader1**: 20 segmentations (.ppm format) of the retinal vasculature from Grader 1
2. **Grader2**: 20 segmentations (.ppm format) of the retinal vasculature from Grader 2
3. **Raw images**: 20 RGB ophthalmic images (.ppm format) of 10 diseased and 10 healthy retinas. Each raw image has a corresponding Grader 1 and Grader 2 segmentation.

Data Preprocessing:

1. Convert the raw RGB retina images to grayscale.
2. Change scale of the segmentations to [0,1] so that black pixels excluded from the vasculature segmentation are 0 and those included as vasculature are labeled 1.

Assumptions:

1. Every raw image has a corresponding Grader1 and Grader2 segmentation
2. The same naming convention applies to all images in the three folders (Raw images, Grader1, and Grader2), namely "im####.[extension]".
3. The "ah" in "ah.ppm" and "vk" in "vk.ppm" are initials of the human Graders and not different file types.
4. The segmentations of each image contain only black [0, 0, 0] and white [255, 255, 255] in RGB space, which becomes [0,255] when converted to grayscale.
5. The images were all collected on the same device by the same individual.

Observations of the data:

1. All images and segmentations are the same size, namely [605×700×3], where 3 indicates the RGB color channels.
2. All images are the same data type, namely uint8.
3. There are 10 images from patients with eye diseases and 10 images from healthy patients.
4. The Kolmogorov-Smirnov (KS) Test tell us that the intensity distributions between the background and the vasculature are all significantly different (using both sets of ground truth segmentations) with $p<0.05$ for all cases. This is encouraging as it suggests it's possible to separate the vasculature from the noisy background.

Data Challenge for a Digital Health Company – Time Limit: 40 Hours
Submitted by: Kalina Slavkova (kslav@sas.upenn.edu)

Visualization of raw images and ground truth segmentations:

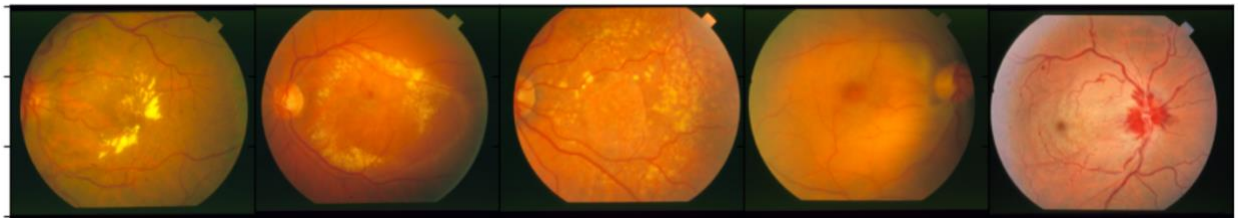


Figure 3. First 5 raw RGB retinal images from the dataset.

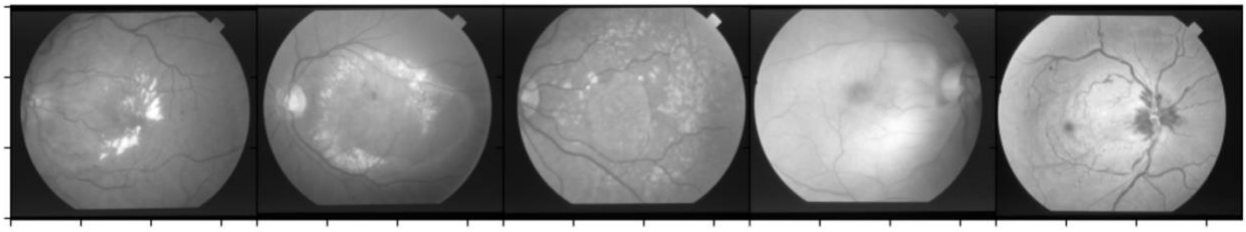


Figure 4. Grayscale transformation of Figure 3.

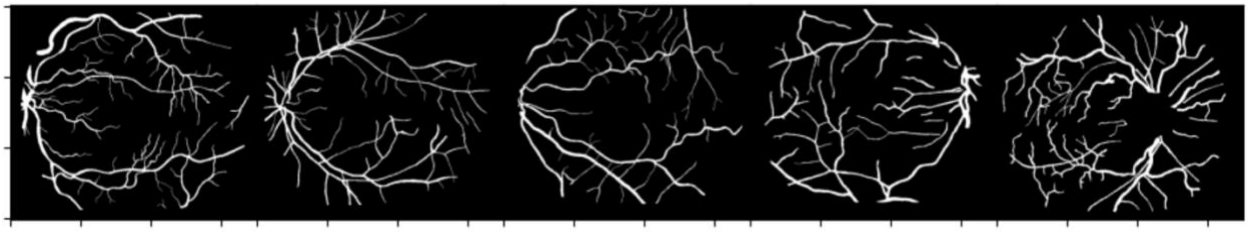


Figure 5. Grader1's segmentations of the first five cases from the dataset from preceding two figures.

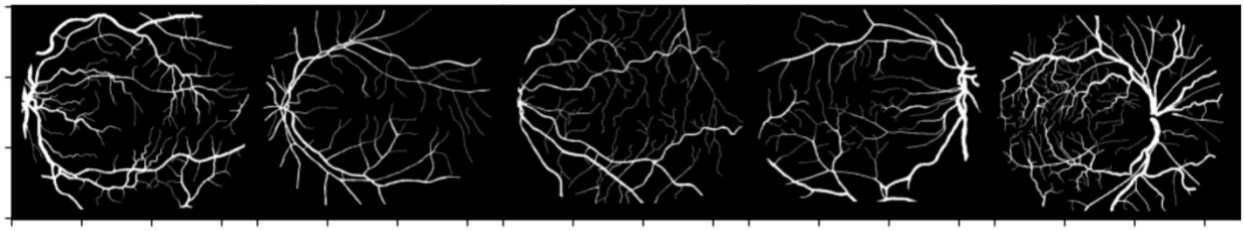
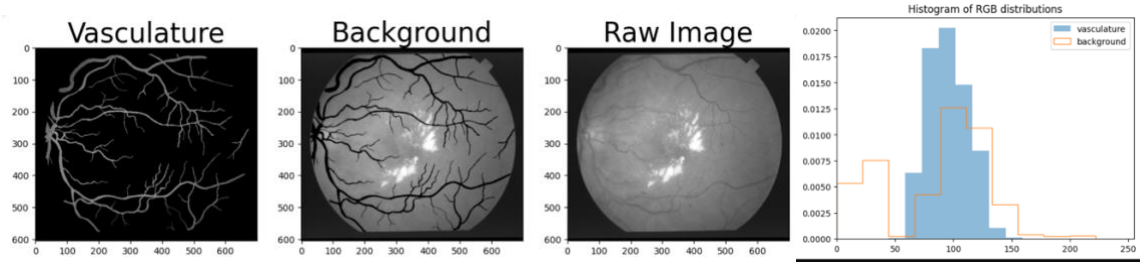


Figure 6. Grader 2's segmentations.

Visualization of the background and vasculature pixel distributions for a case:



For all cases, we find $p<0.05$ from the KS two-sample test.

Data Challenge for a Digital Health Company – Time Limit: 40 Hours
Submitted by: Kalina Slavkova (kslav@sas.upenn.edu)

Segmentation of the vasculature: Apply classical image transformations that will isolate the vasculature from the noisy background.

Since we only have 20 cases, it's not possible to rely on machine learning methods that require training and validation data in large quantities. Instead, we can turn to classical computer vision techniques to manipulate the image in such a way that reduces the background noise to increase contrast between the vasculature and background. These techniques are very fast and should be able to give a good preliminary segmentation.

Working with gray scale images for ease…

1. We first determine a threshold for a mask that will separate the retina from the dark background at the periphery.
2. Next, we apply a Gaussian blur filter to denoise the image and ensure that the vasculature stands out more clearly against the background.
3. Next, we apply adaptive thresholding to determine a local threshold (rather than a global threshold) for a given kernel size, which is useful when the lighting across the image is nonuniform.
4. Finally, we apply a morphological transformation (erosion with dilation for further noise removal) with a rectangular structuring element to select the neighborhood of pixels on which to apply the transformation.

Note: There are many hyperparameters in these steps that can be further tuned for optimizing segmentation performance, but given the time limit of this challenge, optimization was not performed.

Metric: The DICE coefficient will be used as the metric for determining the segmentation performance.

Other options include simple accuracy (number of pixels accurately assigned over the total number of pixels), but this metric is flawed, which is exacerbated when the segmentation of interest is sparse (like vasculature). DICE penalizes incorrect segmentation assignments and is widely used in segmentation tasks. A DICE of 1 indicates perfect segmentation overlap with ground truth, whereas 0 indicates no overlap.

How good is the performance of the proposed model in terms of DICE score with the given hyperparameters?

The highest DICE score among both graders that we observe is **0.749**, while the lowest is **0.406**. Typically for medical applications, we like to see a DICE of 0.70 or above, indicating 70% or higher overlap between the segmentation output and the ground truth reference. The mean scores of **0.6035** and **0.5336** for Graders 1 and 2, respectively, are lower than ideal. All dice scores are stored in the file **dice_[timestamp].csv.**

The human readers are better able to identify and segment minuscule vessels, whereas the proposed algorithm struggles to identify these small vessels from the background noise. On

visual inspection, the retina with artifacts of disease appear to occlude parts of the vasculature, inducing more failure in the proposed segmentations.

There is opportunity for further hyperparameter optimization, like the selected threshold in adaptive thresholding and the size of the kernels in the Gaussian blurring and morphological transformations.

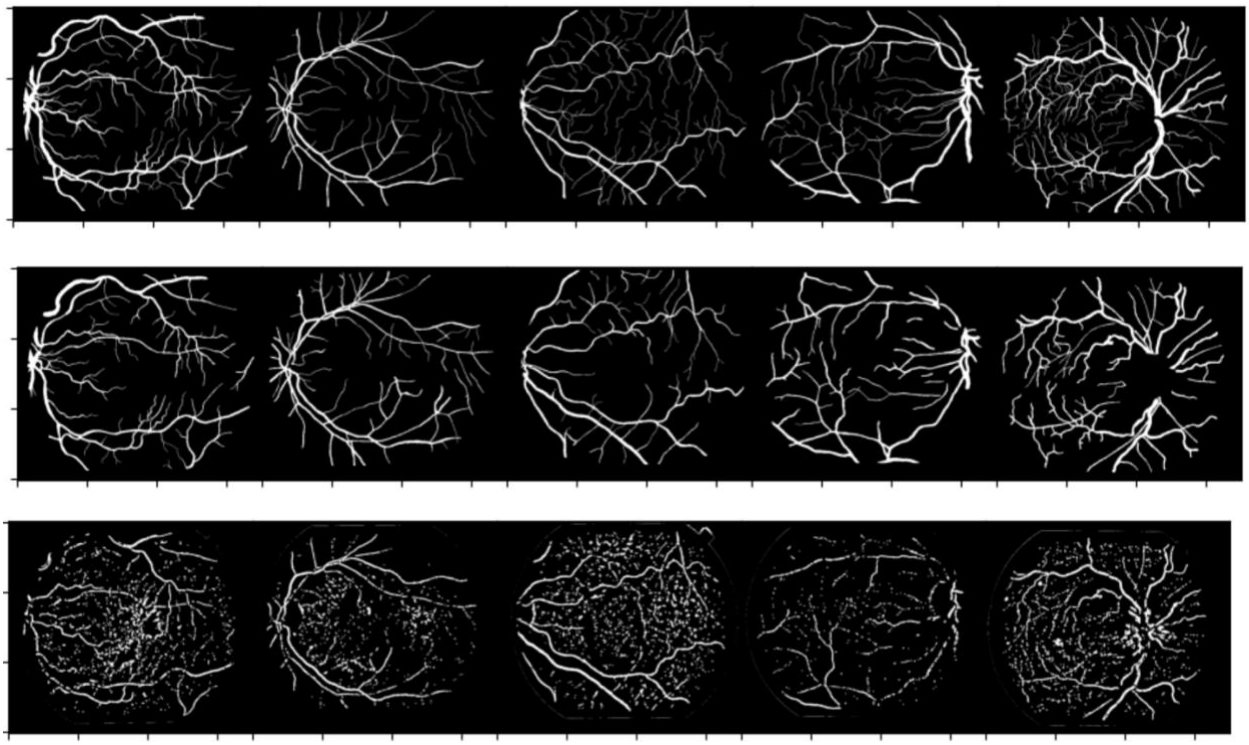


Figure 7. Segmentations from the proposed algorithm for the first 5 cases in the dataset. The first row contains the first 5 segmentations by Grader1, the second row contains the analogous segmentations by Grader2, and the last row contains the analogous segmentations outputted by the proposed algorithm.

ID	Dice (Grader1)	Dice (Grader2)
Im 0001	0.5004673	0.476932
Im 0002	0.5755756	0.569034
Im 0003	0.5374836	0.4767103
Im 0004	0.507562	0.4464696
Im 0005	0.5459735	0.5222416
Im 0044	0.5637934	0.5349608
Im 0077	0.691456	0.6135212
Im 0081	0.6946638	0.5747958
Im 0082	0.6737935	0.5838431
Im 0139	0.5991536	0.5316434
Im 0162	0.6717276	0.5431704
Im 0163	0.7488242	0.6558652
Im 0235	0.6897689	0.5820561
Im 0236	0.6990382	0.6044736
Im 0239	0.5600754	0.4984067

Data Challenge for a Digital Health Company – Time Limit: 40 Hours
Submitted by: Kalina Slavkova (kslav@sas.upenn.edu)

Im 0240	0.6180565	0.5322371
Im 0255	0.6481147	0.549438
Im 0291	0.5549021	0.5194985
Im 0319	0.4594324	0.4064065
Im 0324	0.5311695	0.4496218
Mean:	0.6035	0.5336

What is the image quality of the data?

Because there are no reference images for quality assessment, we can use BRISQUE (Blind/Referenceless Image Spatial Quality Evaluator) as well as a blur score to get an idea of the quality.

BRISQUE scores range from (0,100), with lower scores representing better perceptual quality. The default reference models for comparision, based on natural image scenes, were downloaded from github at https://github.com/opencv/opencv_contrib/tree/master/modules/quality/samples.

The blur score -- computed using the second derivate across the image and hence past rate of change across edges -- tells us whether there are sharp edges. Larger blur scores therefore indicate clearer images and less blur/smear across edges. We compute these two scores for each of the 20 cases and store them in **qualmetrics_[timestamp].csv**.

ID	BRISQUE	Blur
Im 0001	29.4009514	24.7007221
Im 0002	21.9077702	33.7208626
Im 0003	31.0891705	35.9894597
Im 0004	21.2186298	28.0279696
Im 0005	25.7982082	41.1258088
Im 0044	20.9377251	35.4666647
Im 0077	20.9942017	56.539644
Im 0081	20.2790585	50.4197456
Im 0082	19.9792862	56.4759681
Im 0139	26.9462452	37.2719262
Im 0162	15.6689491	65.2146166
Im 0163	21.0902367	57.5889333
Im 0235	19.0118999	51.5242343
Im 0236	19.6954041	51.9677106
Im 0239	18.0374146	68.8920188
Im 0240	23.9439087	37.988579
Im 0255	19.318594	49.7271931
Im 0291	19.2616768	57.5660527
Im 0319	19.7912464	53.8396949
Im 0324	20.3047771	64.0160823

Data Challenge for a Digital Health Company – Time Limit: 40 Hours
Submitted by: Kalina Slavkova (kslav@sas.upenn.edu)

Is the DICE score influenced by the presence of disease?

For determining significance of DICE when stratified by diseased vs. healthy retinas, we can apply the Wilcoxon Rank Sum test. This is appropriate given the small sample size in each category (N=10 in diseased and healthy, respectively).

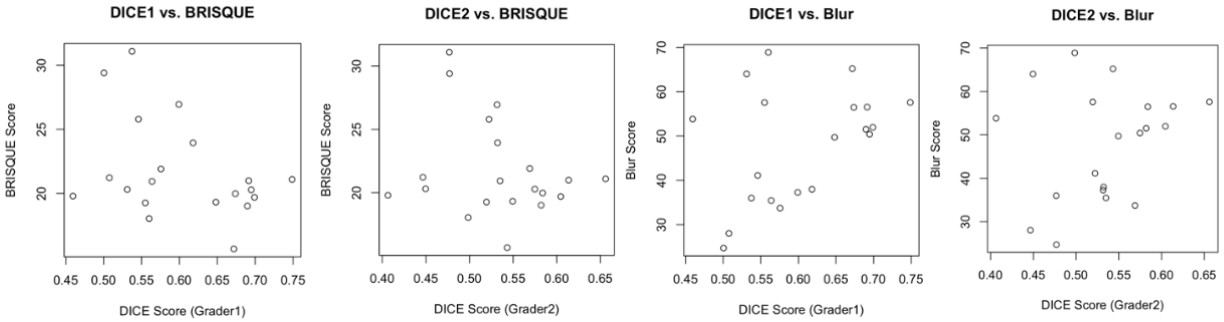
- *DICE (Grader1)*: $p=3.788\times10^{-5} << 0.05$
- *DICE (Grader2)*: $p=5.250\times10^{-4} << 0.05$

There is a statistically significant difference in DICE score when computed for segmentations from healthy retinas versus segmentations from diseased retinas.

	Normal retina	**Diseased retina**
Dice (Grader1) Mean	0.6696	0.53755
Dice (Grader2) Mean	0.5738	0.4934

Is the DICE score influenced by the image quality?

For determining correlations between dice and image quality, we can first plot DICE by both image quality metrics. If the relationship appears linear, we can compute Pearson's correlation coefficient. Else we can compute Spearman's rank correlation coefficient (N=20 is not ideal but workable).



***p*=0.0989** ***p*= 0.3029** ***p*=0.0833** ***p*=0.3391**

It's not very clear what the relationship is. We can compute Pearson's correlation coefficient as a quick test (acknowledging there may be some error if the distributions aren't quite linearly dependent). Again, we're looking for $p<0.05$ as cut-off for significance.

All *p*-values from Pearson's correlation test are above 0.05, so we cannot conclude that there is a statistically significant linear relationship between DICE score and image quality metrics, BRISQUE and Blur.

Are deep learning approaches feasible?

Supervised deep learning algorithms for segmentation are limited by the availability of high quality annotated ground truth data for training.

If we have thousands of annotated examples for training, validation, and testing, then it would certainly be feasible to apply a simple architecture, like a U-Net to this segmentation task. Considerations for preprocessing would include normalization as well as resampling and resizing to ensure all images are the same size and resolution. We would also need to ensure that there are no class imbalances between diseased and healthy eyes so that the network can "see" ample examples of all possible cases.

There are also ample pre-trained networks in the literature that could be fine tuned for this specific task and thus reduce the data requirement; however, one limitation in using pretrained models is the need to preprocess the data for fine tuning in a way that the pretrained model expects for input.

The Medical SAM (Segment Anything Model) is a recent work that shows a lot of promise for various segmentation tasks in medical imaging. The authors have published their code, making it possible to reproduce it!

Proposed SQL application for querying csv file:

I know very basic SQL commands but have not extensively worked with it in my research.

In terminal, to query, one first has to design a database .db file

- sqlite3 database.db

To import a csv file and assign it to the variable “pat_history” from the established database, one can run the following in terminal:

1. .mode csv
2. .import patient_history.csv pat_clinical

Various operations can be performed to extract data from “pat_clinical” with the following format:

- SELECT [insert command here] FROM pat_clinical;

To select the “cigsPerDay” column from pat_clinical, we can use the following command:

- SELECT cigsPerDay FROM pat_clinical

Data Challenge for a Digital Health Company – Time Limit: 40 Hours
Submitted by: Kalina Slavkova (kslav@sas.upenn.edu)

I am confident that if I needed to use SQL consistently in my position, I would be able to pick up the knowledge quickly and apply it. My track record of learning programming languages is strong. For example, I learned R at my Postdoctoral position at Penn (now at Columbia) without ever having worked with it before. Before that I learned MATLAB and Python (PyTorch for deep learning) as a graduate student and published first author works with those toolsets.