Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220441

Multi-Dimensional Laboratory Test Score as a Proxy for Health

Bar H. EZRA [a], Shreyas HAVALDAR [b], Benjamin GLICKSBERG [c], and Nadav RAPPOPORT [a,1]

[a] *Department of Software and Information Systems Engineering, Ben-Gurion University of the Negev, Israel*

[b] *Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York*

[c] *Hasso Plattner Institute for Digital Health, Icahn School of Medicine at Mount Sinai, New York, New York*

Abstract. The standard of care for a physician to review laboratory tests results is to weigh each individual laboratory test result and compare it to against a standard reference range. Such a method of scanning can lead to missing high-level information. Different methods have tried to overcome a part of the problem by creating new types of reference values. This research proposes looking at test scores in a higher dimension space. And using machine learning approach, determine whether a subject has abnormal tests result that, according to current practice, would be defined as valid – and thus indicating a possible disease or illness. To determine health status, we look both at a disease-specific level and disease-independent level, while looking at several different outcomes.

Keywords. Laboratory Tests, UK Biobank, Machine Learning, Electronic Health Records

1. Introduction

An essential part of the clinical decision-making process depends on interpreting different laboratory test scores relative to reference values. This process is usually performed by scanning the results and looking for abnormal values. Reference ranges have sev- eral limitations [1]. One of the core limitations is that the scan for abnormalities is done marginally, where every test is considered independently of the results of the other laboratory tests. Such practice can miss information in higher dimensions where a result of a laboratory test is assessed given the results of other laboratory tests. Another significant issue is that reference values are determined without a specific outcome in consideration, thus potentially creating a false display of abnormal values. Additionally, reference ranges are not always built on representative populations. Existing approaches have tried to cope with the reference values problem by developing different types of multivari- ate reference ranges [2,3]. Previous approaches used laboratory tests from hospitalized patients [4] or relied on longitudinal data [5], or tried to predict long-term effect [6].

[1] Corresponding Author: Nadav Rappoport; E-mail: nadavrap@bgu.ac.il.

We describe two approaches to cope with the limitations described above. The first is a disease-specific method, where the 'normality' of a set of laboratory tests for a given diagnosis is compared between affected and unaffected subjects. The second is a disease-independent method, using Machine Learning methods to detect abnormalities at a higher dimension point-of-view. The goal is to establish a method for alerting the physician that a patient is at a higher risk of having an illness or a potential risk and should be considered for further evaluation.

2. Materials and Methods

2.1. Datasets

Two datasets were used during this research – UK BioBank (UKBB) and data courtesy of the Mount Sinai Data Warehouse (MSDW).

The primary dataset used is the UKBBB which contains data of about 500,000 adult participants from all over the UK. This dataset consists of different types of features, such as demographic features, laboratory test values, previous and current illnesses as well as lifestyle, death information, and more [7]. This dataset was used as a baseline dataset. A total of 33 features including demographic data and laboratory test results were used as features, while disease diagnoses and admission data were used as outcomes of interest. There are known differences between sex in laboratory test results as well as in diseases' prevalence. Therefore, for the Disease-Specific model we split the analysis by sex and describe here females-based ansyslsis (267,746 participants). While for the Disease-Independent model, data from all participants were used. Participants with missingness were excluded, leaving 105,538 participants for the Disease-Specific model and 409,896 for the Disease-Independent model. Data was standardized by reducing the mean and scaling to unit variance by dividing by the standard deviation.

The secondary dataset that was used is the MSDW which aggregates clinical data from five hospitals within the Mount Sinai Health System in New York City. The wellness visits of the Mount Sinai Data Warehouse consists of about 100,000 visits for a wellness check-up. It contains demographic features, laboratory test values, and admission data, where the latter was used as the outcome of interest.

Laboratory tests with over 30% missingness were excluded. We were left with 31 different laboratory test types. Patients with missing data were excluded.

2.2. Disease-Specific Model

In this approach, we looked at a specific illness (e.g., Fatty Liver) and estimated the 'typical combination' of a set of laboratory tests. The objective was to test how well a multi-laboratory test-based score discriminates between ill and healthy patients. Subjects were split into two disjoint groups – ill and controls and were compared based on their scores. The score of subject i was defined as the Euclidian distance of the vector representing i's laboratory test results from the population's average. We hypothesis that this score is associated with illness.

$$Score(i) = d(\text{labs}_i, (labs)) \quad (1)$$

2.3. Disease-Independent Models

In this approach, we wanted to identify a general abnormality in the laboratory test scores by looking at the scores in a higher level of wellness. We trained Random Forest models using the following features age, sex, laboratory test results, and laboratory tests scores as defined above. We didn't use reference ranges to mark abnormal values to stay unbiased. The outcomes of interest represent general health measures as diagnoses and hospitalization rates. Data was split into 70% for models training and 30% for testing. Model's hyperparameter tuning was performed using cross-validation for each model to ensure the best results. Random Forest model was selected as it can capture non-linear effects and is robust to missing data.

We used two types of outcomes of interest: Number of diagnoses and number of admissions post laboratory test. The diagnoses-based outcome was defined according to the number of ICD10-CM codes assigned to each subject. Three binary outcomes were determined according to the number of diagnoses a subject had: at least 1, 5, or 10. A separate Random Forest model was trained for each of these outcomes. Admission-based outcomes were defined within two timeframes: whether readmission was recorded within 30 or 365 days. Due to high rates of imbalance of readmission outcomes, we trained a Balanced Random Forest model.

3. Results

3.1. Disease-Specific Score

We computed the distance of the representing vector of laboratory test results from the average vector for each subject. This distance was used as the subject's score. This analysis was performed for female subjects only. We hypothesized that our defined score is associated with the prevalence of specific diseases. The hypothesis was tested for six different diagnoses: Fatty Liver (ICD10-CM K760), Anemia (ICD10-CM D50*), Neutropenia (ICD10-CM D70*), Parkinson, Asthma, and Alzheimer.

Most subjects (266,511 out of 267,746) had at least one laboratory test outside the reference range. We found the cases and controls for all six diseases have a statistically significant difference in the distribution of the scores based on the Kolmogorov–Smirnov test. All p-values < 0.001 (Figure 1). Moreover, we found clear discrimination between cases and controls based on ROC AUC (Table 1). Moreover, to compare the current practice of marginal tests, we compared the discrimination of our score to the discrimination of other single laboratory tests using ROC AUC. We found that our score outperformed any other single laboratory test discrimination for Fatty Liver disease, Anemia, Neutropenia, and Asthma, but not for Parkinson's and Alzheimer's disease (Table 1). It should also be noted that the mean scores indicate that for Parkinson's and Alzheimer's disease, it is only slightly better than random classification.

3.2. Disease-Independent Score

In this method, we used the UKBB data as discovery data for the diagnoses prevalence model and for readmission rates models. The MSWD data was used for readmission models' validation, as this is the only outcomes information available in this dataset.

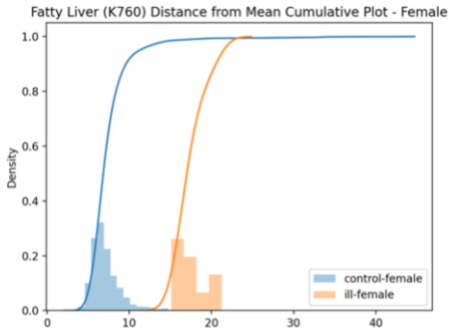

Figure 1. Distribution of the disease-specific scores among cases and controls for Fatty Liver.

Table 1. Comparison between top 5 Laboratory tests' ROC AUC scores and the Disease-Specific test ROC AUC score. GGT: Gamma Glutamyl Transferase, HGB: Hemoglobin, LYM: Lymphocyte, CREAT: Creatinine, EOS: Eosinophill, Glu: Glucose, MCH: Mean Corpuscular Hemoglobin, WBC: White Blood Cell, NEU: Neutrophill, HDL: High Density Lipoprotein, RBC: Red Blood Cell, Alb: Albumin, ALK: Alkaline, HCT: Hematocrit, MONO: Monocyte, LDL: Low Density Lipoprotein, MCV: Mean Corpuscular Volume, PLT: Platelet

Diagnoses	Top 1	Top 2	Top 3	Top 4	Top 5	DS-Score
Fatty Liver	GGT	Glu	HDL	ALK	HGB	**0.700**
	0.618	0.579	0.577	0.568	0.560	
Anemia	HGB	MCH	RBC	HCT %	MCV	**0.650**
	0.617	0.607	0.605	0.601	0.598	
Neutropenia	LYM %	WBC	LYM	NEU	PLT	**0.581**
	0.563	0.559	0.552	0.551	0.542	
Parkinson	CREAT	EOS %	LYM %	LYM	RBC	**0.525**
	0.541	0.533	0.529	0.523	0.515	
Asthma	EOS %	NEU	WBC	MONO	MCH	**0.551**
	0.542	0.516	0.512	0.512	0.511	
Alzheimer	CREAT	Cholesterol	Alb	LDL	LYM	**0.545**
	0.546	0.532	0.530	0.526	0.525	

Table 2. Performance of the three classification models. The prevalence and results are based on the 30% test set.

Observations	Prevalence	Accuracy	ROC AUC
At least 1 Logged Observations	74.3%	0.741	0.64
At least 5 Logged Observations	45.7%	0.633	0.68
At least 10 Logged Observations	26.0%	0.757	0.71

We created three different models for three binary outcomes based on a minimal number of diagnoses, 1, 5, or 10. As can be seen (Table 2), when looking at the ROC AUC score, increasing the threshold for cases definition increases the model's discrimination. We created two different models for each dataset for two different time horizons: readmission within 30 days and readmission within a year. The results from both datasets show that the model performs better when a longer time horizon is used (Table 3).

Table 3. ROC AUC results of the readmission rates model based on the 30% test set.

Timeframe	UKBB	MSDW
30 Days	0.64	0.68
365 Days	0.66	0.72

4. Discussion

In this study, we showed that using the shared information of common laboratory test results is advantageous over marginal information carried by individual laboratory tests. The advantage of a multi-dimensional score includes increased risk for specific diseases like Fatty liver diseases and for general health status as measured by the number of diagnoses and risk for readmission. Although we showed an association, it does not imply any causation. Meaning, we cannot conclude that shifting specific laboratory test results toward the mean or into the reference range will improve the health status. Using the number of diagnoses as an outcome of interest is not ideal, as the diagnoses come from the subject's history and are not limited to diseases diagnosed post laboratory test was taken. This limitation does not apply for readmission, where we limited to outcomes occurring post laboratory test taken. There exists other laboratory tests-based prediction models which are limited to hospital data or rely on longitudinal data [4,5] where the approached described in this work is based on routinely laboratory test results.

5. Conclusions

Considering the multi-dimensional distribution of common laboratory test results may be useful for alerting care providers of potential abnormalities. It can also be beneficial for predicting a specific illness and better or comparable to a single test assessment. The Disease-Independent score models can be used as an alert tool for physicians to indicate patients that may have been overlooked.

References

[1] Rappoport N, Paik H, Oskotsky B, Tor R, Ziv E, Zaitlen N, et al. Comparing ethnicity-specific reference intervals for clinical laboratory tests from EHR data. The journal of applied laboratory medicine. 2018;3(3):366-77.
[2] Boyd JC, Lacher DA. The multivariate reference range: an alternative interpretation of multi-test profiles. Clinical chemistry. 1982;28(2):259-65.
[3] Malka R, Brugnara C, Cialic R, Higgins JM. Non-parametric combined reference regions and prediction of clinical risk. Clinical chemistry. 2020;66(2):363-72.
[4] Froom P, Shimoni Z. Prediction of hospital mortality rates by admission laboratory tests. Clinical chemistry. 2006;52(2):325-8.
[5] Razavian N, Marcus J, Sontag D. Multi-task prediction of disease onsets from longitudinal laboratory tests. In: Machine learning for healthcare conference. PMLR; 2016. p. 73-100.
[6] Horne BD, May HT, Muhlestein JB, Ronnow BS, Lappe´ DL, Renlund DG, et al. Exceptional mor- tality prediction by risk scores from common laboratory tests. The American journal of medicine. 2009;122(6):550-8.
[7] Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS medicine. 2015;12(3):e1001779.