Automated Patient Risk Classification Tool

Applied AI Healthcare Challenge

Huron Consulting Group

1 Introduction

In response to the Applied AI Healthcare Challenge, Huron has used the innovative technologies of robotic process automation (RPA), machine learning (ML), and artificial intelligence (AI) to develop a tool assisting medical professionals to identify patients' potential risk for health conditions. The tool can improve the accuracy of diagnoses, provide clarity and understanding for patients, and fast-track the prescription of treatments.

2 Relevant Questions

2.1 What the Tool Does

Our software uses an ML algorithm with the National Health and Nutrition Examination Survey (NHANES) as the source of data, which is provided by the Centers for Disease Control and Prevention (CDC). The algorithm finds the most impactful questions that define the result of a response variable, in this case a condition a doctor seeks to evaluate in a patient.

The algorithm uses NHANES data to learn how to predict health outcomes of future patients based on past patient responses, through an ML process known as model training. Once this algorithm for a given health condition is trained on the patients in the survey data, the tool asks the provider for information about a patient. The tool then assesses the patient's overall risk for said condition. As the user inputs a patient's information, ML compares it to the general population data, resulting in a statistical probability of the patient having or not having the indicated condition. If the tool believes the patient has an above-average risk, it produces a concise report for the patient to better understand his/her/their condition and treatment. Section 3 provides a more detailed technical overview of the process.

2.2 Today's Practices and Limitations

Today, a doctor diagnoses patients based on the symptoms the patients explain, testing they may perform in the office, and their own knowledge of medicine. Often, patients may be referred to multiple specialists, and even after testing there can be levels of uncertainty.

Misdiagnoses — defined as missed opportunities to make a timely or correct diagnosis based upon available evidence — affect more than 5% of Americans per year, or around 12 million adults [https://qualitysafety.bmj.com/content/23/9/727.]. One study estimates that over 100,000 Americans die or are permanently disabled each year because of incorrect or delayed diagnoses [The 'Big 3' in Diagnostic Errors (hopkinsmedicine.org)]. Our demo focuses on the diagnosis of cancer, which according to the same study accounts for almost 38% of misdiagnoses in the US each year. Doctors play a crucial societal role to monitor health conditions and trends in the general population. By understanding health outcomes and trends in the community, health providers can more quickly and accurately provide treatment plans. Other than formal reports and broader research articles that investigate specific health conditions, the current market lacks sufficient analytics-based tools readily available for healthcare providers to quickly compare a patient's health information against the general population. The absence of a diagnostic tool like this creates a risk for the patient, who may be subject to a provider's inherent research bias when further investigating their condition.

2.3 Innovating to Promote Success

Many misdiagnoses can be attributed to cognitive bias, which is estimated to account for up to 70% of diagnostic errors [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6982612/#CR1.] This model helps avoid errors caused by unavoidable human biases by using ML, which objectively comes to conclusions based on vast data sets. Our innovative algorithm classifies patients through comparison with data from the general US population. The algorithm can evaluate whether an individual may be at an elevated risk for a specific condition. By providing this quick, accurate, and robust risk classification, this tool enables providers to make more effective, informed, and accurate diagnoses and prescriptions of treatment.

2.4 Making a Difference

Upon full deployment, this tool can help eliminate delayed diagnoses and mitigate the risk of cognitive bias influencing a diagnosis, resulting in more accurate and predictable results. This model saves doctors time when making a diagnosis. It adds an additional tool to their arsenals to help confirm their initial diagnoses by identifying an objective and data-supported assessment of patient risk. Automated patient risk classification could potentially save thousands of lives each year by increasing providers' speed and precision. As a result, hospitals and medical centers could save millions of dollars each year by decreasing malpractice lawsuits, saving hospital resources from treating misdiagnoses, and reducing the overall administrative workload of the clinical staff. This tool could also give patients direct access to an easy-to-use resource helping to determine the priority of seeing a medical professional in the first place.

2.5 Implementation Risks and Payoffs

The adoption curve associated with AI in the healthcare sector poses a potential risk. The tool's users must understand how and when it should be deployed. Although the implementation of this model is quick and easy, a medical professional must still make an official diagnosis. This tool is intended to augment other diagnoses steps and tests for confirmation. It does not replace evidence-based practices or established testing methods. While any novel AI model poses some degree of skepticism, patients can rest assured knowing that the diagnosis is vetted for accuracy by a qualified clinician and that their healthcare isn't solely in the hands of a single machine or human. This method combines the quality assurance of provider expertise and experience with the precision and vast power of analytics available through RPA, AI, and ML.

The potential payoff of this tool's deployment is a risk classification tool aiding the diagnostic process of providers investigating health conditions and increasing awareness of risk factors.

2.6 AI Alignment & Safety Engineering

Our demo uses attended automation (i.e., it runs only when initiated by the user, in this case a licensed provider). Beyond the original source data, which in this demo comes from NHANES, our model only has access to data input by a medical professional. In our demo, the dataset provided by the CDC has already been de-identified; this process uses no Personally Identifiable Information (PII). The technologies used in our tool all store data locally; the tool won't require any external sources (i.e., cloud connections) and is completely encapsulated inside the system it gets loaded into. Once the process has run its course, the result is given to the patient or doctor in the form of an HTML file that isn't stored within the model after transmission. The HTML file is an isolated process preventing the risk of any information being accessed by an outside party, beyond what is input over the course of usage.

3 In-Depth Process

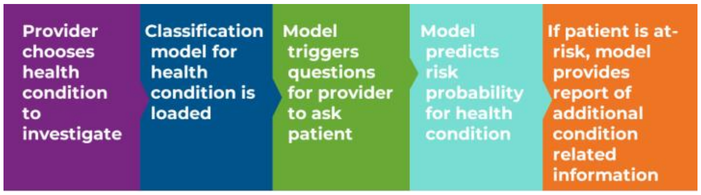
Provider chooses health condition to investigate
Classification model for health condition is loaded
Model triggers questions for provider to ask patient
Model predicts risk probability for health condition
If patient is at-risk, model provides report of additional condition related information

Figure 1: High-Level Business Process. By leveraging the power of RPA, ML, and AI technologies, our model eliminates delays and reduces risk to improve the accuracy and efficiency of patient risk classification.

3.1 Input Data

Our model can be trained to work with numerous types of data sets, but we chose data from NHANES because it is extensive and publicly accessible. NHANES surveys a representative sample of the US population and data from the survey is used by many researchers to draw conclusions on the health of the general public. Once the data has been read into our tool from this data source, a column from the data set must be chosen to tie to the health condition we are looking to investigate, called the response variable. This column can be seen as an "answer" on which our classification algorithm bases its prediction. In our demo, our response variable is whether a person has been told in the past they have cancer.

3.2 Elastic Net Regression - Feature Selection

Currently, there are likely to be hundreds of variables in the dataset of our model, all containing different information about a patient. The question we are looking to answer is, "which of these variables are most important for predicting the health condition we are investigating?" The answer leverages the power of ML and more specifically, an Elastic-Net Regression Model.

Regression models are commonly used in ML to approximate the relationship of a response variable (in our case, whether a patient has cancer), to the rest of the features in a dataset. Elastic-Net Regression Models go one step further in their functionality to increase the importance of more predictive features while reducing the significance of less important features in our overall model. This allows us to perform feature selection, which reduces the number of input variables in our overall classification model. Instead of asking the provider to fill out hundreds of columns of data pertaining to a given patient, feature selection provides a statistical method allowing us to only identify the most important features, which we can choose to pass into our classification model.

3.3 Logistic Regression - Classification

We now have a filled dataset containing the most significant features that strongly correlate to the health condition we are investigating. We are then able to train a classification model, using only these selected variables, over the rows of patient data that we have from the cleaned NHANES dataset. This classification model uses a technique called logistic regression, another type of regression model that we can now use to make predictions on the overall risk that a new patient may have to a given health condition. This type of regression model lends itself extremely well to classifying binary response variables (yes/no answers), just like in the case of our cancer response variable, and the response variables of other health conditions we wish to investigate in this survey.

3.4 Overall Risk Assessment

Using the logistic regression model, we can compare the new patient's overall risk of the health condition to the average overall risk of the general population found within the NHANES dataset. If the patient is found to be at an elevated risk by our tool, it provides the patient with a report containing supplemental information about the health condition. This report is manually created and stored before the running of this tool. Each report is tailor-made to each health condition, with information such as symptoms, potential treatments, and additional recommendations included for the patient to consider regarding next steps.

4 Production/Deployment Considerations

4.1 Tool Advantages

Flexible Inputs - This tool is currently based on the NHANES dataset, but it can be easily adjusted for other sources of patient data and increase in accuracy if the dataset it connects to is more comprehensive.

End-User Adaptability – While our demo envisions a medical professional as the end-user, this tool can also be adjusted to be patient-facing, with the patient screening themselves before even seeing a provider.

Friendly User Interface (UI) and Quick Deployment - One of the main benefits of using RPA is it allows this ML model to be easily deployable as an out-of-the-box package on an end user's local machine with a friendly UI and easy navigation not requiring any analytics expertise.

Multiple Health Conditions - Although this demo focuses on the solution's capabilities to predict cancer risk, this tool can be built out for multiple health conditions at the user's discretion.

4.2 Technology Requirements

Our model uses three distinct types of innovative technology – RPA, ML, and AI. Firstly, we implement an RPA software allowing users to code "bots" able to automatically perform repetitive, routine tasks, triggered either by human intervention, the time of day, an email, etc. It can recognize buttons, images, and text, interact with browsers and any desktop application (e.g., Excel), and process faster than any human. Next, we wrote an ML algorithm with Python, a popular coding language for data science. This algorithm can make conclusions about millions of data points and efficiently analyze a multitude of variables to determine what's relevant or not. These two technologies combine to create an AI solution that both navigates a given user's machine and performs decision-making to classify health risks.

5 Federal Agency Use Cases

We have identified a sample of example use cases where the risk classification tool has the potential to thrive at different federal agencies. This is in no way a comprehensive list of possible use cases.

5.1 Centers for Disease Prevention and Control (CDC)

The dataset we are using is publicly hosted by the CDC at: *https://www.cdc.gov/nchs/nhanes/index.htm.* We could deploy this tool as a dashboard addition to this website with this model's process as the base. Any user accessing the website could go to the webpage that holds this risk classification tool, enter information about themselves, observe their overall risk to various health conditions, and then decide if they wish to see a provider.

5.2 Veteran's Health Administration (VHA)

The VHA is one federal healthcare system that could benefit greatly from actively using this tool. VHA could give on-site medical providers the ability to quickly compare incoming patients' information with that of the general population. By making this tool provider-facing, it can be used as an additional aid to determine if a patient should be tested further to confirm a potential health risk and decrease the system's overall misdiagnosis rate.