Matt Ferguson

End-of-Life Advisor Machine Learning Ethics

Objective

We are tasked with addressing ethical considerations which result from a contract to create an "end-of-life-advisor" machine learning model. The contract value is financially significant and may mean the difference between business success and failure. Many patients with terminal pathology have a "living will" that will give consent for care to be withdrawn in cases of no reasonable hope for recovery The end-of-life advisor will assess whether treatment should be withheld from patients with terminal pathology. If the model was to determine there is no reasonable hope for recovery, doctors may decide to withdraw care resulting in patient death. Therefore, we have a duty to examine the ethical issues which result from implementation of an end-of-life machine learning model.

Medical Ethics

Before a direct examination of the ethical issues raised by end-of-life machine learning models can begin, we must establish a framework of ethics. Classical medical ethics can be summarized with the Hippocratic oath's key phrase "first, do no harm". On the surface this is a straightforward tenant. We administer only treatments which either help or at least do not harm the patient. However, a deeper inspection of this phrase and classical medical ethics in general will reveal its failure to meet implications linked with modern treatments.

Imagine a cancer drug which cures cancer in 99% of terminal cancer cases yet is wholly fatal in 1% of cases. From an unlucky individual's point of view, this drug would cause grave harm. We do not have the ability to 100% accurately predict which individual patient will be harmed by a treatment. In fact, many drugs for individuals with severe disease have worse harm reduction performance than our toy example. This hypothetical drug causes harm objectively in limited instances.

I argue a pharmaceutical administered by a medical professional which causes *any* harm objectively violates the Hippocratic oath in a literal sense and for this reason classical medical ethics fall short. Imagine the millions of people who would be doomed to terminal cancer if we implement this strict interpretation of the Hippocratic oath and avoid all treatments that are not 100% safe. We would be reducing our ability to cause harm to zero, while also minimizing our reduction of harm so dramatically that we would set modern medicine backwards hundreds if not thousands of years. The Hippocratic oath states "do no harm" in plain and clear terms. However, we can interpret this text in a non-literal sense as we do with other philosophical and religious texts to gain greater insight.

We can repurpose this classical medical ethic from a utilitarian perspective by employing statistics. All or most medical interactions of the time were at the one-on-one level, and the effectiveness of available medicine was minimal. Hippocrates created the Hippocratic oath in the same century as Euclid discovered geometry. Mathematics and science were less advanced than today. The field of statistics did not exist (and while I believe the great Greeks had an intuitive understanding of statistics; it was the Arab mathematicians during the Islamic golden age who first formulated the field). With greater technical abilities we can now evaluate the term 'harm' with statistical scrutiny.

Let us set forth a definition of "do no harm" in a statistical formation where all patient outcomes are measured. The net benefit of a treatment across a patient population would be the 'net harm reduction'. If this net harm reduction is positive and significant, then the pharmaceutical would do

no ‘net harm’ and would greatly benefit the human species. With statistical operations, we have left behind an operation on individuals and arrived at operations at the species level. A cold statement.

The way harm is measured in this statistical formulation must be objective and consistent with reasonable human expectations. Imagine if we were to measure outcomes as binary. In this case, a death would be weighed as much as a stubbed toe. If the pharmaceutical’s sole benefit is to cure baldness and we measured its net harm reduction using binary terms it could kill 1 patient and cure the baldness of 2 patients and maintain a net harm reduction score. The true harm of a treatment must be measured in continuous terms and the magnitude of a harm should be commensurate with its impact on the human experience.

Many doctors are no longer working with single patients. They work on pharmaceuticals or tools that may benefit millions of people. The individual variations in individual’s body chemistry and unique pathology makes it impossible for any single tool or treatment to have a uniform outcome. With the greater efficiency gained by worker specialization (i.e., pharmaceutical researcher), and corporatization (i.e., pharmaceutical company) many modern medical interactions if not all operate on a statistical level rather than a one-on-one level. A statistical or utilitarian perspective of net harm reduction will serve best in modern medicine.

Ethical Considerations

With this utilitarian formation of medical ethics set forth we can continue to answer the question posed. We consider that we are a founding partner at a company which has been tasked with developing a “end-of-live advisor” machine learning model. This model would use statistical features to predict when a patient has become terminal and untreatable. Terminal and untreatable together mean that the patient will die soon irrespective of treatment. There is a real difference between the ground truth of terminal/untreatable and a doctor’s or model’s prediction of such condition. In that gulf between our best prediction and the unknowable reality lie the deepest ethical dilemmas.

How will this data be collected, and how will bias be controlled for? Let us imagine that we collect all traditionally available measurements and attributes of patients at hospitals and mark the ground truth as to whether care was withdrawn. We have already introduced potential bias into our dataset. The dataset may have demographics that are not wholly consistent with other hospitals. A model based on an average hospital may have reduced performance in hospitals with abnormal demographics. A model trained on regular patients will not perform well at a children’s hospital or a hospice for the elderly. The length of stay of types of patients will be much longer than others, which will be another source of bias in the dataset. A fatal car crash will have particularly short length of stay, versus say a slow growing tumor from prostate cancer. These examples will be used for training, and we would need to pay close attention to how our model would perform on rare instances or rare demographics.

Are there any privacy or security concerns relating to an end-of-life advisor? In the medical field, the Health Insurance Portability and Accountability Act (HIPAA) stipulates how healthcare information and personally identifiable information must be protected from fraud and theft. Hospitals and other healthcare providers are prohibited from disclosing any protected medical information including predictions made by an end-of-life advisor for an individual patient. Obviously training records must be acquired within the framework provided by HIPAA and they must be sanitized of personally

identifiable information. The predictions made by our machine learning model would be protected by HIPAA, though it must be possible for decisions made by our model to be audited while maintaining HIPAA protections. Regarding security, modern cybersecurity practices must be employed to prevent any healthcare data or personal data from being leaked.

Which statistical feature are to be used in this model? Those with and without traditional health insurance may have different levels of care, and different treatment outcomes. I think we can certainly set out that health insurance status should not be an input to an end-of-life advisor! We are prohibited from discriminating based on race, age, sex, and other intrinsic characteristics in many contexts. In business and legal settings these safeguards against discrimination are the norm. These prohibitions stem from western liberal principles and egalitarianism. However, do they make sense in the context of medical treatment?

Age especially is an extremely meaningful statistical indicator of the effectiveness of treatment. Objectively, 80-year-old cancer patients are more terminal than 20-year-old cancer patients. However, is a 5-year-old cancer patient more terminal than a 20-year-old cancer patient? Likely so. And if that is the case, then what is the ethical path forward then? Deny care to a child that has a lesser likelihood of treatment in favor of giving resources to a young adult with a better chance of survival? A harrowing question. A model trained on age will make decisions regarding the extension of treatment to the very young and elderly which are ethically fraught. However, age must be an input to the machine learning model. Without age we would be greatly degrading the performance of our model. We leave room for ethics by yielding to doctors the ultimate say in whether treatment is extended or not. There must be human oversight of an end-of-life advisor for human moral predispositions to be respected, especially with respect to moral end-of-life decisions made for the very young and elderly.

In the recent post-pandemic era, we can evaluate ethical issues which resulted from race as an input to healthcare. There was notably higher mortality and rates of infection among Hispanic and African Americans. These groups were hospitalized for COVID-19 twice as often as Whites. A contributor to this racial discrepancy may be the varying obesity rates between Whites, Hispanics, and African Americans. Dr. Daniel C. DeSimone M.D. of the Mayo Clinic adds that racism, type of work (such as service industry jobs), location (such as an urban dwelling), and access to healthcare are the culprits. Frankly, I believe there are other reasons for this discrepancy we can examine but we must press on to ethical considerations. Should race be an input to an end-of-life advisor? If certain ethnic groups were to be more likely to be affected by certain terminal disease, would this make the model more likely to recommend that care be withdrawn? I suspect the extra events will make the end-of-life model more accurate for those ethnic groups. However, if the correlation of one ethnicity to our target variable terminal/ untreatable is correlated more than another we have another ethical dilemma. The patient's ethnicity would be at least a part of the model's rationale for withdrawing or extending care. I suspect most disease will not have a perfect correlation across race, and that race is at the very least a statistically insignificant predictor of many diseases. We can narrow our scope of consideration only to those diseases where race is statistically significant. If race is a statistically significant predictor, then should it be a model feature? Two perspectives exist, one which states race is a proxy of other attributes, and one which states race is a meaningful taxonomy. In the society where race is simply a proxy of other attributes, those other attributes must be measured and given to

the model. In the society where race is taxonomically meaningful (traits vary across race, and race is not just skin color) then race must be an input to the model.

The chief ethical concern is whether a machine learning model should make life and death recommendations. Let us evaluate the types of errors a machine learning model may make in this situation to better understand the danger of this type of machine learning deployment. The model in our simplistic case would output two decisions, terminate care, or extend care. This would give us two prediction errors, false treatment extensions, and false treatment terminations. The cost of a false treatment extension prediction is financially significant but more ethically sound. We extended care to an individual who would not benefit from care. The reverse situation occurs for a false treatment termination. An individual who may have recovered from terminal illness had their care withdrawn and now faces death. We must actively engineer against false treatment terminations as this poses an extreme ethical dilemma. I think for the mentioned reasons our model output must not be binary. It must be multiclass if not fully continuous. Cases which are near class boundaries may disproportionately affect outcomes. Imagine if we were 49% likely to recover and the model output a 0 (terminate care). A doctor would look much more favorably on a 0.49 vs a 0. A smaller dilemma arises if we extend care to an individual who would not benefit from care, and thereby deny care from another. For certain types of diseases there is a true scarcity of treatment. The relative scarcity of treatment may be an input to the model. On one hand, the terminality of disease will be a fundamental input to the model. Treatment follows the same laws of economics as the rest of our world. However, adding the scarcity of treatment as an input to our model will only serve to make it more restrictive. In cases where there is plentiful treatment it will extend treatment as normal, but for those rare treatments our model would grow much more restrictive and not reflect the true medical reality but instead a 'medical-economic' reality. Our utilitarian ethic reflects learning gained from the fields of statistics and economics and perhaps a model that understands scarcity of treatment can be truly helpful during a pandemic situation in maximizing outcomes. Further research and ethical consideration should be given in this area. However, our scope of consideration is to accurately predict end-of-life and not adjust our predictions by the relative availability of treatments. Therefore, scarcity of treatment should not be a consideration for our model.

We have created a set of well supported arguments to answer the ethical concerns of an end-of-life advisor. First being, medical ethics are best served by a utilitarian formation which maximizes reduction of harm by statistical means. Close attention must be given to training and deployment demographics to prevent biasing model performance. Age and race are statistically predictive of treatment outcome, however it is for society to weigh the value of these as inputs to healthcare. Room must be left for the moral weight we place on the young and elderly and this may be accomplished with human supervision over deployed models. Finally, false treatment termination is a grave prediction error which must be engineered against using machine learning best practices.

References

1. Greek Medicine - The Hippocratic Oath (nih.gov)
2. Health Insurance Portability and Accountability Act - Wikipedia
3. COVID-19 infections by race: What's behind the health disparities? - Mayo Clinic