Can Artificial Intelligence help reduce human medical errors? Two examples from ICUs in the US and Peru

Miguel Paredes[1]

[1]MIT

February 19, 2018

Abstract

Medical human-errors cost society billions of dollars in the US and around the world. A recent estimate claims that measurable medical errors in the US accrued to US$17.8 billion dollars in 2008. Many of these errors can be prevented by applying current artificial intelligence techniques, e.g. machine learning, on existing medical data in order to predict these high-risk error situations. Since most of the settings leading to human-induced errors are known and well-studied, predicting and preventing them is more attainable. Using US and Peruvian ICU data, we develop, demonstrate, and discuss how machine learning models could possibly aid in identifying and preventing medical human-errors. We propose that artificial intelligence is most valuable in helping prevent or reduce medical human-errors when doctors must make decisions where clinical trials are lacking, providing some guidance and support to doctors grounded in historical data.

1. Introduction

Medical human errors cost society billions of dollars in the US and around the world (Van Den Bos et al., 2011). Many of these human errors go undetected, making it difficult to have a clear understanding of the magnitud of the problem. Therefore, estimates of the negative effects of human errors on society have been derived through many different methods (Thomas and Petersen, 2003). The types of errors that harm patients receive the name of measurable medical human errors, and provide a lower bound or conservative estimate of the actual effects of these types of errors on society.

The economic effects of these measurable errors translate into both direct and indirect costs. Direct costs are mainly due to an increase in the medical costs of providing inpatient, outpatient, and prescription drug services to individuals who are affected by medical errors. Indirect costs are related to increased mortality rates among individuals who experience medical errors and related to lost productivity due to related short-term disability (Chmieleski et al., 2010).

Many studies have tried to quantify the direct and indirect costs of measurable medical human errors (Brady et al., 2009). The most recent group of studies places the accrued costs of measurable human-erros in the US at US$17.8 billion dollars in 2008 (Van Den Bos et al., 2011), with 2,500 excess deaths and over 10 million excess days missed from work due to short-term disability (Chmieleski et al., 2010), and an estimate of 1 in 4 injury related visits to hospitals in the US were subject to medical errors between 2008 and 2009 (David et al., 2013). While these studies focus on estimating the cost of measurable medical errors, usually extrapolating findings from one dataset or clincal domain to a more general population, most studies do not focus on actions to reduce the heavy costs society must bear due to these errors.

Our study calls attention to what can be done to help reduce medical errors, using as ilustratory examples AI models developed based on ICU data from the US and Peru. Two research questions motivate our investigation: 1) What types of medical errors are best address through AI models, and 2) how can AI help doctors and other medical professionals make better decisions. Specifically, we explore how AI-based models can help predict what patients should or should not receive diuretics when in septic shock in US ICUs, and we explore how AI-based models can help doctors decide what children to admit into an ICU in Peru. We hypothesis that by using machine learning - a subfield of AI - in high-risk decision settings we could help see a reduction in medical human errors.

The article is structured as follows. Section 2 provides a literature review on medical human-errors, artificial intelligence, and medical applications. Section 3 presents our experimental setup as a means to explore our proposal of how AI could help the medical space, describes the data and AI-based prediction models. Section 4 presents and discusses the mortality prediction results for the ICU data in the US. This paper ends with conclusions, policy implications, and next steps in Section 5.

2. Literature review

Multiple studies have tried to answer the question of what are the costs of medical human errors. One of the first and most influential studies, due to its methodological and statistical sampling rigor, was the Harvard Medical Practice Study (HMPS), which found that adverse events were a common component of hospital care (Brennan et al., 1991; Leape et al., 1991). The HPMS study employed a two stage chart review methodology in which nurses first analyze patient records with high likelihood adverse event presence, and then doctors thouroughly review selected charts to confirm possible adverse events and to evaluate the ocurrance of suboptimal care. The HMPS determined 1984 incidence rates for all types of medical injuries in New York, estimating healthcare costs of US$3.8 billion, implying national cost of errors slightly above US$50 billion (Johnson et al., 1992). While large population-based chart reviews put forth by the HPMS methodology are not without criticism, they are still widely used and have recently been validated and used at a national level (Thomas and Petersen, 2003).

A study focusing on medical injuries in Utah and Colorado found that the total costs for preventable medical errors in the two states in 1992 came up to US$308 million (in 1996 dollars), implying national costs of errors of about US$17 billion. The study was based on the review of medical records from a representative random sample of 14,732 discharges from 28 hostpitals in 1992 (Thomas et al., 1999). In 1999, the Institute of Medicine released a report titled *To Err is Human,* estimates that 98,000 Americans die any given year in hospitals due to medical human-errors based on extrapolations from the HMPS and the Utah-Colorado study (Donaldson et al., 2000). A more recent study found that the rate of medical errors was 133.3 per 1,000 hospitalizations, with affected patients incurring 18.5 percent more in hospital charges and having a 14.6 percent longer hospital stays than patients not exposed to medical errors (Layde et al., 2005).

One of the most recent studies estimated medical error costs at about US$19.5 billion in the United States during the year 2008 (Van Den Bos et al., 2011). Most of this increased cost (US$17 billion) was due additional services needed by individuals affected by these medical errors. The study also finds indirect costs increases due to higher mortality rates (US$1.4 billion) and lost productivity levels brough on by short-term disability (US$1.1 billion). The study uses medical claim data for a large insured populationto the United States population, and by extrapolating to the United States obtaines obtains 6.3 million medical injuries in 2008, out of which the authors estimate that 1.5 million were associated with a medical error. The study found that the total cost per error was approximately $13,000, obtaining a total cost of US$19.5 billion in medical error in the US. Additionally, these errors representes 2,500 additional deaths and more than 10 million additional days missed from work due to short-term disability (Chmieleski et al., 2010).

The most common and costly types of measurable medical errors in the US in 2008 were postoperative infection (US$3.4 billion), pressure ulcer (US$3.3 billion), mechanical complication of noncardiac device, implant or graft (US$1.1 billion), and postlaminectomy synfrom (US$1 billion), accounting for almost half of all estimated medical error associates costs that year (Van Den Bos et al., 2011). In their Utah and

Colorado study, Thomas et al grouped adverse events into five large categories of errors: operative, drug related, diagnostic or therapeutic, procedure related, and other, based on individual chart reviews. Of these errors, complications post-operations were the most costly (39% of all medical errors) (Thomas et al., 1999).

Medical errors in the Intensive Care Unit (ICU)

Intensive care units (ICUs) are hospital departments in which patients who are dangerously ill are kept under constant observation and given intense care. Given the critical nature of patients in ICUs, human medical errors can have even more severe adverse effects than in other hospital departments.

The causes for medical errors have been studied and documented (Donchin et al., 1995). One prospective observational study of 391 found 120 adverse events in 79 patients (20.2%), including 66 (55%) nonpreventable and 54 (45%) preventable adverse events as well as 223 serious errors, which occurred during the ordering or execution of treatments, especially medications (61%; 170/277) (Rothschild et al., 2005). The study found that while many types of errors were identified, the most common type of error was to carry out the intended treatment. Another prospective study in a pediatric ICU identified 52 medication errors throughout 26 12-hr observation periods, which included 357 reviewed written orders and 263 observed doses . Of the 52 medication errors, 42 (81%) were considered clinically important (Buckley et al., 2007). One study found that sleep deprivation was a major determinant of human medical errors, specially among medical interns, and that reducing the number of hours interns work per week can reduce serious medical errors in the ICUs (Landrigan et al., 2004).

Can AI help address medical errors in ICUs

In this study, we seek to explore whether AI can help reduce human medical errors. The term "artificial intelligence" (AI) is not easy to define, reason why many different definitions exist. The word "artificial" refers to homething that doesn't occur naturally, while the word "intelligence" has been defined in many ways. The psychologist Howard Gardner proposes a definition that focuses on problem-solving: "Intelligence is the ability to solve problems, or to create products, that are valued within one or more cultural settings" (Howard, 1983). Society sometimes classifies as "artificial intelligence" only those activities that it perceives as "hard" for computers to do (like correctly describing what is occuring in an image) in contrast to "simpler" taks computers more often do today (like calculations in spreadsheet).

An intuitive definition of AI proposed by MIT Professor Thomas Malone suggests that AI is "machines acting in ways that seem intelligent" (Malone, 2017). A more formal definition proposed by MIT Professor Patrick Winston's says that "AI is about the architectures that deploy methods enabled by constraints exposed by representations that support models of thinking, perception, and action" (Winston, 1984).

There are two main categories in the field of AI: "narrow AI" and "general AI". Narrow AI can be defined as "a machine-based system designed to address a specific problem" (Kiron, 2017), while general AI refers to machines that can solve a diverse array of types of problems on their own, similar to humans. Currently, all known applications of AI are narrow AI types. While general AI is one of the most active research topics today, experts such as Ray Kurzwel and Patrick Winston predict that general AI applications are decades away (Creighton, 2018).

Diverse studies have also shown that artificial intelligence and machine learning algorithms can help doctors make better decisions, many times outperforming their human counterparts on the diagnosis of certain illnesses or in the prediction of certain medical outcomes such as mortality or length of stay (Saly et al., 2017; Ghassemi et al., 2014; Pirracchio et al., 2015; Henry et al., 2015; Mayaud et al., 2013). One type of medical error is one when doctors make a wrong prediction regarding the appropriate treatment for ICU patients. Paredes et al explore this in the context of US ICUs and conclude that machine learning could aid physicians by providing better predictions about the effect of certain treatments and the likely evolution of sepsis patients (Paredes and O'Reilly, 2016).

Despite the promises and hype around AI, there has been much criticism and unfullfiled promises as well. One example is Watson, IBM's AI software that has been marketed heavily as the ultimate physician companion, but has yet to deliver any results (Ross and Swetlit, 2017).

3. Experimental Setup

The task of employing AI to help reduce medical human-errors in the medical space is not easy nor straighforward. In order to envision these benefits, many of which will take years to materialize, we propose two experiments using medical data and machine learning to demonstrate how AI can help improve decision making, and hopefully reduce human medical errors. We draw from Paredes et al for the work on US ICUs and complement this work with an analysis of ICU data from Peru.

The purpose of these experiments is to demonstrate how AI, and in particular machine learning, can help predict health outcomes, thus having the potential to support decisions in the ICU. While we do not predict medical errors through our models, we demonstrate the predictive power of these models in order to inspire and demostrate how prediction high-risk medical error situations could be possible.

In our first experiment, we employ AI to help doctors better understand if sepsis patients should receive diuretics by predicting the effect of diuretics on two medical outcome (mortality and length of stay). In our second experiment, we employ AI to predict child mortality, thus potentially helping doctors understand what patients to admit.

3.1. Data

We employ data from intensive care units (ICUs) from the US and Peru. For the US, we employ the MIMIC dataset, which is a large database containing information relating to patients admitted to critical care units at a large tertiary care hospital in Boston. MIMIC includes vital signs, medications, laboratory measurements, observations and notes charted by care providers, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data, and more (Johnson et al., 2016). For Peru, we employ data from Peru's Children's Hospital (PCH), which has been collected in the last decade by Dr. Tantalean, one of PCH's former directors.

US ICU Experiment - MIMIC database

We pre-process the MIMIC database for our analysis. After data extraction and filtering we obtain a study group of 1,522 patients who were or had been in septic shock (3.81% of the entire MIMIC database). Out of these 1,522 patients, 189 received diuretics (D^+)(12.4% of study group), and 1,333 (87.6% of study group) did not receive diuretics (D^-). For each patient, age, gender, and race were obtained along with 21 clinical features as can be seen in table 1. These 21 clinical features were measured on three days (day of entry, day before and day of diuretics decision) for a total of 66 features per patient. Additionally, there are two outcome measures for each patient: 30 day mortality after ICU discharge, and ICU length of stay.

Table 1 Patient Features MIMIC database

Age
Gender
Race (white / non-white)
SAPS II Score
SOFA Score
Elixhauser Comorbidity Index
Congestive Heart Failure Indicator
Cardiac Arrhythmias Indicator
Valvular Disease Indicator
Hypertension Indicator
Uncomplicated Diabetes Indicator
Complicated Diabetes Indicator
Renal Failure Indicator
Liver Disease Indicator
Obesity Indicator
Creatinine Level (day of measurement average)
Fluid Inputs (day of measurement sum)
Fluid Outputs (day of measurment sum)
Vasopressor Use
Mechanical Ventilation Use
Maximum Blood Pressure (on measurement day)
Average BP (on measurement day)
Mortality (did patient die within 30 days of leaving ICU)
Length of Stay (days)

A fundamental problem for doctors treating sepsis patients in ICUs is that the patients arrive with varying degrees of illness severity, so what worked for one patient in the past might not work for a very similar patient. Additionally, there is no standard protocols for determining if a patient should receive diuretics, which is one of the main treatment alternatives for septic patients. Thus, an AI model that helps predict whether a patient will benefit from diuretics taking into account all available data could help doctors make better decisions and reduce errors related to treatments.

Table 2 summarizes the average values for all clinical variables within each group for D^+ and D^- patients on the day of ICU entrance and the day of diuretics decision, and shows whether these mean differences are statistically different through a Welch two sample t-test.

Table 2 Cohort Balance between D^+ and D^- patients. Summary of average feature values. Systematic differences between D^+ and D^- patients

Variable Name	D^+	D^-	**Different Distribution?**
Age (Years)	66.24	66.14	No (0.93)
Gender (% of males)	42.86%	42.46%	No (0.92)
Race (% of whites)	**0.53%**	**2.18%**	**Yes (0.01)**
Day of ICU entrance SAPS II Score	**15.78**	**14.99**	**Yes (0.02)**
Day of Diuretics Decision SAPS II Score	**17.93**	**17.06**	**Yes (0.03)**
Day of ICU entrance SOFA Score	**9.52**	**7.37**	**Yes (0.00)**
Day of Diuretics Decision SOFA Score	**10.35**	**8.94**	**Yes (0.00)**
Elixhauser Comorbidity Index	3.164	2.993	No (0.20)
% with Congestive Heart Failure	**48.15%**	**31.36%**	**Yes (0.00)**
% with cardiac arrhythmias	**37.04%**	**25.13%**	**Yes (0.00)**
% with valvular disease	**13.76%**	**8.25%**	**Yes (0.00)**
% with hypertension	26.98%	27.46%	No (0.89)
% with uncomplicated diabetes	24.87%	19.35%	No (0.10)
% with complicated diabetes	4.76%	5.78%	No (0.55)
% with renal failure	5.29%	8.63%	No (0.07)
% with liver disease	8.47%	9.38%	No (0.68)
% with obesity	2.65%	1.35%	No (0.29)
Day of ICU entrance Creatinine	**1.570**	**1.88**	**Yes (0.01)**
Day of Diuretics Decision Creatinine	1.623	1.85	No (0.06)
Day of ICU entrance Fluid Inputs	1010.86	1113.63	No (0.26)
Day of Diuretics Decision Fluid Inputs	**3089.24**	**2560.45**	**Yes (0.03)**
Day of ICU entrance Fluid Outputs	**1994.13**	**1385.59**	**Yes (0.00)**
Day of Diuretics Decision Fluid Outputs	1652.45	1596.06	No (0.82)
Vasopressor (% of patients)	**0.86**	**0.65**	**Yes (0.00)**
Mechanical Ventilation (% of patients)	**0.94**	**0.68**	**Yes (0.00)**
Day of ICU entrance Maximum BP	114.31	114.57	No (0.87)
Day of Diuretics Decision Maximum BP	108.32	110.91	No (0.09)
Day of ICU entrance Average BP	**77.12**	**79.23**	**Yes (0.04)**
Day of Diuretics Decision Average BP	**74.85**	**78.09**	**Yes (0.00)**
Mortality (% that died)	0.33	0.378	No (0.17)
Length of Stay (days)	**15.18**	**6.30**	**Yes (0.00)**

From Table 2 we can see that diuretics patients have significantly different health conditions upon entering the ICU (different distributions equal to Yes), and also at the moment when the decision to receive diuretics was taken. This suggests that we face selection bias. Patients who receive diuretics are more ill when they come into the ICU (higher SAPS II, SOFA scores), and on the diuretics decision day (SAPS II score, SOFA score). This worse health can be a confounding factor by influencing diuretics decision and health outcomes. Additionally, diuretics patients seem to have overall higher morbidity (congestive heart failure, cardiac arrythmia, and valvular disease). We also observe higher fluid levels, higher creatinine levels, and higher need for mechanical ventilation or vasopressors for D^+ patients.

Given this systematic difference, we will need to employ a model that addresses selection and confounding bias, and that can potentially help doctors predict whether diuretics (or another course of treatment) wil be effective.

Peru ICU Experiment - PCH database

We pre-process the PCH database for our analysis. After data extraction and filtering we obtain a study group of 1,708 patients. Out of these patients, 321 patients do not survive (S^-) (18.7% of study group), and 1387 survive (S^+) (81.3% of study group). For each patient, age, gender, weight, height and clinical characteristics were obtained for a total of 32 features as can be seen in table 3. Additionally, we have one outcome measure for each patient (30 day mortality), which will be the variable we are trying to predict.

Table 3 Patient Features PCH database

Age
Gender
Proceeding department
Proceeding medical service
Reason for admission
Traqueotomy
Glucose level upon entry
Mechanical Ventilation Use
Post operatory entry
Congenial malformation
Nutritional level
Weight
Height
Number of failing organs
Respiratory organs deficiency
Cartiovascular organs deficiency
Neurological organs deficiency
Hepatic organs deficiency
Septic
SIRS
Catheter
Mortality

3.2. Models

US ICU Diuretics and Sepsis Mortality Model

Our first model uses the MIMIC critical care database to predict the effect of diuretics administration on patients with a sepsis diagnosis by matching diuretics positive patients (D^+) to diuretics negative patients (D^-) using propensity matching (Rosenbaum and Rubin, 1983) following Paredes et al (Paredes and O'Reilly, 2016), but using the updated MIMIC III database and modifying the main method with machine learning as described below. Before matching patients, we observe that D^+ and D^-patients experience the same mortality rates, and that D^+ patients spend 8.8 additional days in the ICU. Moreover, significant health differences between (D^+) and (D^+-) patients are observed on a number of features upon entrance to the ICU. Therefore, and to address the systematic differences between (D^+) and (D^+-) patients we apply propensity score matching (PSM) to control for these differences (Dehejia and Wahba, 2002).

PSM builds a similarity-of-treatment score (i.e. the propensity score or PS) from a set of patient features (D^+ and D^-) and then matches patients on this score. A PS is the conditional probability of assignment to a particular treatment given a vector of observed covariates. In our model, the study group compares treatment (being administered diuretics) vs. no treatment (not receiving diuretics), denoted by the variable z with values 1 and 0. Each patient is represented by a set of covariates $\mathbf{x} = \{x_1, x_2, ..., x_{66}\}$. The propensity score then is the conditional probability that a patient with vector $\mathbf{x}$ of observed covariates will be assigned to treatment, given by:

$$e(\mathbf{x}) = Pr(z = 1|\mathbf{x}).$$

(1)

A PS can be estimated using a logit model for

$$e(x) = \frac{e(y)}{1 - e(y)} = \alpha + \beta^T f(x),$$

(2)

where $y = log[\frac{e(x)}{1-e(x)}]$, α and β are parameters, and $f(\cdot)$ defines a regression function. 66 covariates of the study group were included in the initial propensity score model. Many propensity score models can be built depending on the vector of features defined and the model specification used. Given this ample design space, one common approach is to iteratively explore what combination of features provides a model that maximizes some the gain of information on the outcome of interest.

We develop several diuretics propensity models by running logistic regressions of the different feature sets on the diuretics indicator. We then use these models to obtain each patient's PS score, which can be interpreted as the probability that a particular patient will be administered diuretics, conditional on the selected health indicators. After the PS is produced for each patient, we discard D^- patients with minimal feature support. For every D^+ patient, we match them to at least one D^-patient based on the similarity measure, and check for experimental group balance to assess how similar they are in terms of health indicators. Finally, we estimate the average treatment effects on the treated (ATT) obtaining standard errors and p-values, by regressing the outcome variable (30 day mortality or length of stay) on one or both of the PS and feature sets.

We use the dataset described in section 3.1 and choose the parameters we will vary. We select five health indicators sets (all features, best feature set defined through a stepwise algorithm, two sets of features suggested by physicians, and a set of features selected through a genetic algorithm). We allow more than one D^- patients to be matched to each D^+ patient, specify that matching can be done with replacement (D^- patients can be matched more than once), define a similarity measure (Euclidean, Mahalanobis, and PS), and select nearest neighbor as the matching algorithm. This combination of parameter definitions leads us to 9 different models for each outcome.

Peru ICU Mortality Model

Our second model compares a machine learning model to PRISM, one of the standard prediction scores used in ICU settings (Pollack et al., 1988). After testing different machine learning models to predict mortality (random forest, support vector machines, naive bayes, CART), we employ a XGBM machine learning model (Chen et al., 2015). We follow the common set up for training a machine learning model where we use 80% of the data to train the model, and then test the model on the remaining 20% (test or hold-out set) and evaluate the performance of the model based on how well it predicts if a patient will survive or not.

4. Results

For our US ICU experiment, we run 9 experiments for each outcome tuning different model parameters, and find that diuretics is associated with a 10% to 18% mortality decrease (compared to similar mortality rates without controlling for differences) and between 5.9 to 8.4 additional ICU days, suggesting that the effect of diuretics was being underestimated when comparing these outcomes but not controling for confounding and selection bias through matching and machine learning. Our machine learning model predicts with 78% accuracy the likelihood that a sepsis patient will die after 30-days of ICU discharge. These results suggest that AI-based support systems could help physicians decide what treatment to provide patients, taking into account vasts amounts of data and finding patters that doctors would normally miss, specially in high-stress environment.

These types of AI-models are especially useful when physicians are faced with decisions for which clear evidence is lacking. In the medical world, evidence is established through clinical trials. However, there are many illnesses or health situations where clinical trials have not been done, and will most likely never happen. It is in these medical situations where machine learning models can help make sense of historical data finding patters and making predictions and recomendations to doctors.

For our Peru ICU experiment, we observe that PRISM predicts 169 survivors (compared to the 261 observed survivers) and 173 deaths (compared to the observed 81 deaths). Thus, current mortality scoring measures would heavily overestimate the risk of children seeking admitence to the ICU. Given that physicians face scarce resources (beds, nurses, medicines, etc.), and that they would like to accept patients whom they can help, the results of PRISM could be leading PCH ICU doctors to accept patients who are not critical, thus restricting access to resources. When we observe the predictions of our AI model, we see that the the AI-based model predicts 230 survivors (compared to the 261 observed survivors) and 111 deaths (compared to the observed 81 deaths). Compared to the PRISM predictions, our AI model is more acurate at classifying patients who will survive and not survive. This increased prediction accuracy could help doctors better assign resources and support their decision of whom to admit into the ICU.

		Predicted		
	Died	No	Yes	Total
Observed	No	140	121	261
	Yes	29	52	81
	Total	169	173	342

Table 1: PRISM Mortality Predictions on the Test set

		Predicted		
	Died	No	Yes	Total
Observed	No	199	62	261
	Yes	31	49	80
	Total	230	111	341

Table 2: AI-based model Mortality Predictions on the Test set

5. Conclusions, policy implications and next steps

Medical errors bring heavy costs on society, both in monetary and non-monetary forms. We proposed that AI can help physicians and other medical professionals in situations where these preventable medical errors are more likely. We show how an AI-based model can help doctors more accurately predict mortality of sepsis patients where they to prescribe diuretics. We also show how another AI model can outperform one of the standard prediction scores used extensively in the medical field.

We believe that physicians are far away from being replaceble by machines. However, we do believe there are many instances in which AI-based models can outperform humans given their ability to process vast amounts of data at incredible speeds, their virtual perfect memory, and their ability to not be affected by human emotions or fatigue. Echoing Obermeyer et al, in the end the application of AI in medicine is likely to be like a team sport, with physicians, nurses, and technical staff supporting themselves heavily on AI-based systems (Obermeyer and Lee, 2017).

However, we recognize that unless data is available, the benefits of AI and machine learning will be non-existent, reason why digital transformation efforts in the medical space are of utmost importance. The two example models presented in this study are basically useless in most medical settings due to the inexistence of similar data.

Next steps include examining a medical claims database in Peru, in order to estimate the costs of medical errors in Peru, similar to the studies done in the US and presented above.

References

Anne Brady, Richard Redmond, Elizabeth Curtis, Sandra Fleming, Paul Keenan, ANNE MALONE, and Fintan Sheerin. Adverse events in health care: a literature review. *Journal of nursing management*, 17 (2):155–164, 2009.

Troyen A Brennan, Lucian L Leape, Nan M Laird, Liesi Hebert, A Russell Localio, Ann G Lawthers, Joseph P Newhouse, Paul C Weiler, and Howard H Hiatt. Incidence of adverse events and negligence in hospitalized patients: results of the Harvard Medical Practice Study I. *New England journal of medicine*, 324(6):370–376, 1991.

Mitchell S Buckley, Brian L Erstad, Brian J Kopp, Andreas A Theodorou, and Gail Priestley. Direct observation approach for detecting medication errors and adverse drug events in a pediatric intensive care unit. *Pediatric critical care medicine*, 8(2):145–152, 2007.

Tianqi Chen, Tong He, Michael Benesty, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, pages 1–4, 2015.

Susan Chmieleski, Michael Dekker, Barbara Scott, Steven M Shapiro, Steve Siegel, Allen Elstein, Derek Jones, Rick Kelly, Sujata Sanghvi, Carl Taylor, et al. The Economic Measurement of Medical Errors. 2010.

Jolene Creighton. The “Father of Artificial Intelligence” Says Singularity Is 30 Years Away. Technical report, 2018.

Guy David, Candace L Gunnarsson, Heidi C Waters, Ruslan Horblyuk, and Harold S Kaplan. Economic measurement of medical errors using a hospital claims database. *Value in Health*, 16(2):305–310, 2013.

Rajeev H Dehejia and Sadek Wahba. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, 84(1):151–161, 2002.

Molla S Donaldson, Janet M Corrigan, Linda T Kohn, et al. *To err is human: building a safer health system*, volume 6. National Academies Press, 2000.

Yoel Donchin, Daniel Gopher, Miriam Olin, Yehuda Badihi, Michal RNB Biesky, Charles L Sprung, Ruven Pizov, and Shamay Cotev. A look into the nature and causes of human errors in the intensive care unit. *Critical care medicine*, 23(2):294–300, 1995.

Mohammad M Ghassemi, Stefan E Richter, Ifeoma M Eche, Tszyi W Chen, John Danziger, and Leo A Celi. A data-driven approach to optimized medication dosing: a focus on heparin. *Intensive care medicine*, 40 (9):1332–1339, 2014.

Katharine E Henry, David N Hager, Peter J Pronovost, and Suchi Saria. A targeted real-time early warning score (TREWScore) for septic shock. *Science translational medicine*, 7(299):299ra122–299ra122, 2015.

Gardner Howard. Frames of mind: The theory of multiple intelligences. *NY: Basics*, 1983.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

William G Johnson, Troyen A Brennan, Joseph P Newhouse, Lucian L Leape, Ann G Lawthers, Howard H Hiatt, and Paul C Weiler. The economic consequences of medical injuries: implications for a no-fault insurance plan. *JAMA*, 267(18):2487–2492, 1992.

David Kiron. What Managers Need to Know About Artificial Intelligence. *Sloan Management Review*, January, 2017.

Christopher P Landrigan, Jeffrey M Rothschild, John W Cronin, Rainu Kaushal, Elisabeth Burdick, Joel T Katz, Craig M Lilly, Peter H Stone, Steven W Lockley, David W Bates, et al. Effect of reducing interns' work hours on serious medical errors in intensive care units. *New England Journal of Medicine*, 351(18): 1838–1848, 2004.

Peter M Layde, Linda N Meurer, Clare Guse, John R Meurer, Hongyan Yang, Prakash Laud, Evelyn M Kuhn, Karen J Brasel, and Stephen W Hargarten. Medical injury identification using hospital discharge data. 2005.

Lucian L Leape, Troyen A Brennan, Nan Laird, Ann G Lawthers, A Russell Localio, Benjamin A Barnes, Liesi Hebert, Joseph P Newhouse, Paul C Weiler, and Howard Hiatt. The nature of adverse events in hospitalized patients: results of the Harvard Medical Practice Study II. *New England journal of medicine*, 324(6):377–384, 1991.

Thomas Malone. Introduction Video - MIT AI MOOC. Technical report, 2017.

Louis Mayaud, Peggy S Lai, Gari D Clifford, Lionel Tarassenko, Leo Anthony G Celi, and Djillali Annane. Dynamic data during hypotensive episode improves mortality predictions among patients with sepsis and hypotension. *Critical care medicine*, 41(4):954, 2013.

Ziad Obermeyer and Thomas H Lee. lost in Thought—The Limits of the Human Mind and the Future of Medicine. *New England Journal of Medicine*, 377(13):1209–1211, 2017.

Hemberg Paredes and Una-May O'Reilly. On the Challenges of using Propensity Score Matching to study Intensive Care Unit patients. *https://arxiv.org/submit/2167333*, 2016.

Romain Pirracchio, Maya L Petersen, Marco Carone, Matthieu Resche Rigon, Sylvie Chevret, and Mark J van der Laan. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. *The Lancet Respiratory Medicine*, 3(1):42–52, 2015.

Murray M Pollack, Urs E Ruttimann, and Pamela R Getson. Pediatric risk of mortality (PRISM) score. *Critical care medicine*, 16(11):1110–1116, 1988.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Casey Ross and Ike Swetlit. IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close. Technical report, 2017.

Jeffrey M Rothschild, Christopher P Landrigan, John W Cronin, Rainu Kaushal, Steven W Lockley, Elisabeth Burdick, Peter H Stone, Craig M Lilly, Joel T Katz, Charles A Czeisler, et al. The Critical Care Safety Study: The incidence and nature of adverse events and serious medical errors in intensive care. *Critical care medicine*, 33(8):1694–1700, 2005.

Danielle Saly, Alina Yang, Corey Triebwasser, Janice Oh, Qisi Sun, Jeffrey Testani, Chirag R Parikh, Joshua Bia, Aditya Biswas, Chess Stetson, et al. Approaches to predicting outcomes in patients with acute kidney injury. *PloS one*, 12(1):e0169305, 2017.

Eric J Thomas and Laura A Petersen. Measuring errors and adverse events in health care. *Journal of general internal medicine*, 18(1):61–67, 2003.

Eric J Thomas, David M Studdert, Joseph P Newhouse, Brett IW Zbar, K Mason Howard, Elliott J Williams, and Troyen A Brennan. Costs of medical injuries in Utah and Colorado. *Inquiry*, pages 255–264, 1999.

Jill Van Den Bos, Karan Rustagi, Travis Gray, Michael Halford, Eva Ziemkiewicz, and Jonathan Shreve. The $17.1 billion problem: the annual cost of measurable medical errors. *Health Affairs*, 30(4):596–603, 2011.

Patrick Henry Winston. Artificial intelligence. 1984.