EHR Safari: Data is Contextual

William Boag WBOAG@MIT.EDU
MIT CSAIL
Cambride, MA, USA

Mercy Oladipo MOLADIPO@MIT.EDU
MIT CSAIL
Cambride, MA, USA

Peter Szolovits PSZ@MIT.EDU
MIT CSAIL
Cambride, MA, USA

Abstract

In the last decade, machine learning (ML) has shown tremendous success in areas such as vision, language, strategic games, and more. Parallel to this, hospitals' capacity for data collection has greatly increased with the adoption and continuing maturation of electronic health records (EHRs). The result of these trends has been a large degree of excitement and optimism about how ML will revolutionize healthcare once researchers get access to data. In this work, we present a cautionary tale of the instinct some computer scientists have to "let the data speak for itself." Using a popular, public EHR dataset as a case study, we demonstrate numerous examples where a non-clinician's intuition may lead to incorrect – and potentially harmful – modeling assumptions. We explore both non-obvious quirks in the data (i.e., hypothetical incorrect assumptions) and examples of published papers that misunderstood the data generating process (i.e., actual incorrect assumptions). This case study is meant to serve as a cautionary tale to encourage every data scientist to approach their projects with the humility to know what they can do well and what they cannot. Without the guidance of stakeholders that understand the data generating process, data scientists run the risk of "garbage-in, garbage-out" analysis because their models are not measuring meaningful relationships.

1. Introduction

From recommender systems to financial trading, data science has transformed the way many sectors opertate in recent years. In 2012, the success of AlexNet demonstrated that deep learning can be incredibly successful at difficult tasks such as object recognition in images (Krizhevsky et al., 2012). One of the main "draws" of deep learning is the ability for models to automatically learn meaningful representations of the input data without explicit feature-engineering (Bengio et al., 2014). The ability to learn without explicit domain knowledge is seen as a tremendous strength of deep learning, allowing it to outperform previous models in computer vision, speech recognition, and more. However, it is always possible to lean too heavily into a paradigm, embracing the ethos of previous generations of Computer Scientists, such as Frederick Jelinek who said, "Every time I fire a linguist, the performance of the speech recognizer goes up." In this paper, we argue that such an orientation is harmful,

especially for high-stakes environments such as healthcare. ML requires an understanding of the data generating process, and its use in healthcare contexts requires computer scientists must work with domain experts like doctors and nurses.

Machine learning is a very broad tool with many possible uses; pattern recognition allows for models to discover even un-intuitive relationships. As a result, ML is used not just for identifying cats in pictures (Le et al., 2012) and sentiment analysis of movie reviews (Pang et al., 2002) but also to predict whether a patient is likely to die (Parikh et al., 2019). Often times the real outcome itself is not even measurable, and proxies are used instead, such as predicting whether a defendant is likely to be re-arrested as a proxy for whether they will re-commit a crime (Skeem and Lowenkamp, 2016) or their level of healthcare costs as a proxy for how sick they are (Obermeyer et al., 2019). ML is so flexible that there are no *technical* limits on what researchers can try to learn associations between, even if social and ethical norms sometimes constrain what should be done.[1]

1.1. Generalizable Insights about Machine Learning in the Context of Healthcare

In this paper, we examine a case study of electronic health record (EHR) data. Our contributions are twofold:

1. We do a deep dive on a popular, public EHR dataset to demonstrate the need for a domain expert to help navigate assumptions about the unintuitive data generating process.

2. We identify some examples of published works which contain incorrect assumptions about the data generating process.

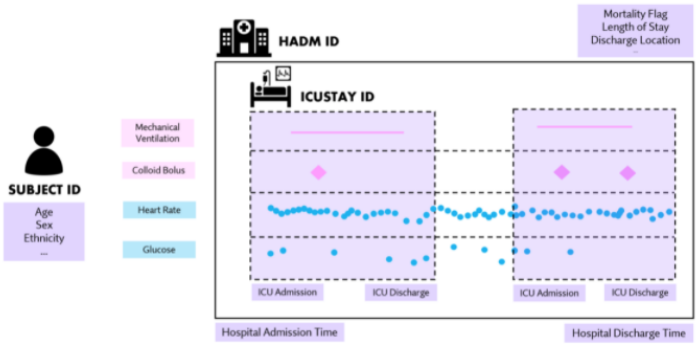
Our intended audience for this work is anyone – particularly technical scientists – who would feel confident in conducting data science and drawing conclusions from data even when they are not entirely sure what the data means. Our hope is that this case study illustrates the pitfalls in common sense assumptions and the need to work closely with a domain expert to understand the data generating process.

2. Background and Related Work

Numerous works have investigated the unintuitive ways that machine learning models have failed. Ribeiro et al. (2016) demonstrate a computer vision model which is 'right for the wrong reason;' using an explanation tool, they show the model was correctly predicting an image was a wolf not because of the canine's features but because the image had snow in the background. Buolamwini and Gebru (2018) demonstrate that commercial facial recognition software performed worse against women and dark-skinned subjects. In the healthcare space, Obermeyer et al. (2019) identify that a model deployed and making predictions for millions of patients had been underrestimating Black patients' risks by using a biased proxy as its training label. Zech et al. (2018) identify that models were achieving inflated results because of confounders, such as a model identifying indicators of which hospital system

1. https://blog.neurips.cc/2021/08/23/neurips-2021-ethics-guidelines

Figure 1: An illustration of the data in MIMIC captured by the EHR, including some data which is ICU-only and some from throughout the hospital stay. A given patient (subject-id) might have multiple hospitall admissions (one per hadm-id). A given hospital admission might have multiple trips to the ICU (one per icustay-id).



(essentially a more sick one or a less sick one) was performing the radiology scan as a proxy for the severity of the image's illness.

In the book *Artificial Unintelligence*, Broussard (2019) argues that because "all data is dirty ... [and that] in every seemingly orderly column of numbers, there is noise," data scientists often have to make many assumptions about the structure of the data. Yang and Roberts (2021) observe that politically-motivated selective censorship of datasets can bias models which are supposed to be general-purpose tools. D'Ignazio and Klein (2020) introduce *Data Feminism*, a framework for critiquing and doing data science using concepts from the academic discipline of intersectional feminism. Data Feminism enumerates seven principles, including:

- **Elevate Emotion and Embodiment**: "[V]alue multiple forms of knowledge, including the knowledge that comes from people as living, feeling bodies in the world."
- **Consider Context**: "[D]ata are not neutral or objective. They are the products of unequal social relations, and this context is essential for conducting accurate, ethical analysis."
- **Make Labor Visible**: "The work of data science, like all work in the world, is the work of many hands. Data feminism makes this labor visible so that it can be recognized and valued."

To mitigate this issue of decontextualization, Gebru et al. (2018) introduce *Datasheets for Datasets* to provide an approach to standardize the process for documenting the data generating process. Extending *Datasheets* to the domain of clinical collaborations, Saleh et al. (2020) introduce *Clinical Collabsheets* to facilitate effective collaboration between data scientists and clinicians. They interviewed 18 experts in the Clinical ML field and distilled the lessons learned into a list of questions to promote productive discussion when working on a new project. Their work presents 8 sets of questions to discuss with an interdisciplinary team at the beginning of a clinical collaboration, including an entire set dedicated to the data

generating process. It encourages the team to discuss: missing data patterns, how/when the data is recorded (e.g., automatically vs nurse input at the end of the shift), label validity, gotchas (e.g., "timestamps reflect time of entry instead of time of event; low acuity (controls) patients all receiving imaging from the same machine; survey-based questions delivered differently by site"), confounders, dataset shift, and curation/pre-processing. .

Within the field of ML for Health, MIMIC-III has been one of the most widely-used public databases (Johnson et al., 2016). Recognizing the importance of curating meaningful representations of patient data, Wang et al. (2020) create *MIMIC Extract*, a preprocessed version of MIMIC, in order to be more useful to researchers for machine learning tasks. This pre-processed dataset successfully standardizes measurement units and performs outlier detection to eliminate physiologically impossible measurements based on expert-curated measurement ranges (Harutyunyan et al., 2019). Additionally, it aggregates similar measurements where appropriate (e.g., in the EHR, "heart rate" is encoded as ItemID 211 before 2008 and as ItemID 220045 from 2008 onwards, due to a change of software systems from Carevue to Metavision). Because of this many of these pre-processing are necessary but unintuitive, *MIMIC Extract* dataset is a valuable contribution to the research community. However, it has been focused on researcher utility rather than focusing on how MIMIC itself can be a case study to demonstrate the challenges of unintuitive modeling assumptions to a larger audience.

3. Data

We use the MIMIC-III v1.4 database (Johnson et al., 2016). MIMIC-III is a large, freely-available database consisting of de-identified EHR data from over 58,000 hospital admissions for nearly 38,600 adult patients. The data was collected from Beth Israel Deaconess Medical Center from 2001–2012. During that period, Beth Israel switched EHR vendors from CareVue to MetaVision. Every hospital admission in MIMIC has at least one ICU stay, though some admissions may have multiple stays. Figure 1 demonstrates some of the data collected during one hospital admission for a patient, including demographics, patient outcomes (e.g., mortality, length of stay, discharge location), laboratory test results, vital sign measurements, medications, procedures, and caregiver notes.

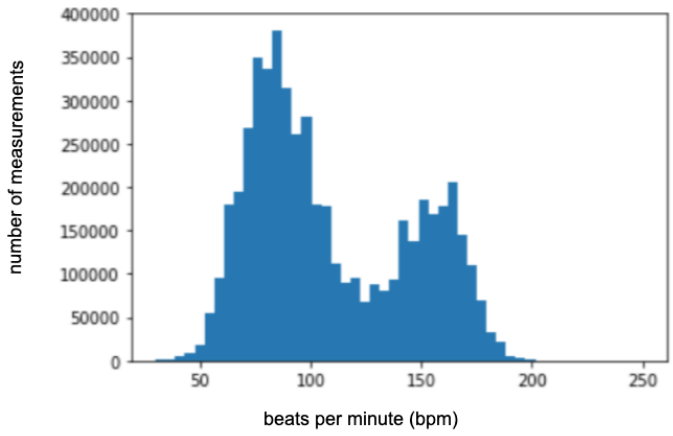
The documentation[2] for MIMIC-III outlines the most important tables, schemas, and inputs and outputs for understanding the data generating process. MIMIC-III has been cited over 3,400 times. As one of the earliest large EHR-based dataset, it has been very influential to the field of ML for Health.

4. Methodology

In this work, we illustrate the pitfalls of making common sense assumptions about domain-specialized data. This is done through two sets of experiments: first, we show a series of counterintuitive "quirks" of a popular EHR dataset, and second we identify two published papers which made incorrect modeling assumptions that undercut the points they hope to make.

2. https://mimic.mit.edu/docs/iii

Figure 2: When plotting heart rate measurements in MIMIC, there is a surprising finding: some patients have hearts that beat 3 times per second.



To examine the EHR quirks, we employ techniques from exploratory data analysis (Tukey, 1977). Professionalized by Tukey in 1977, exploratory data analysis critiqued the overuse of statistical hypothesis testing ("confirmatory data analysis") by the statistics community. As intended, the purpose of this technique is to suggest hypotheses about the causes of interesting phenomena, identify shortcomings in the data to supplement through additional collection, and more.

In Section 6, we critique currently-published research. To accomplish this, we first introduce a given work's motivation, summarize its main findings, and then show how incorrect modeling assumptions undermine the paper's main results. Further, we discuss the implications that these oversights have on the overall findings and impact of the works. This is done with mixed methods, employing both internal analysis (i.e., how the experiment relates to the paper's findings) and external analysis (e.g., literature review of how the work's findings are being referenced by other researchers).

5. Quirks of an EHR Dataset

In this section, we examine surprising artifacts from MIMIC-III's data generating process.

5.1. Surprising Distribution of Heart Rates

According to the Cleveland Clinic, a normal resting adult heart rate should fall within the range of 60 - 100 beats per minute (bpm). However, when we plot the distribution of heart rates vs. age, we see a different story. Figure 2 shows that the distribution of patients in MIMIC have heart rates not only in the normal 60–100 bpm range, but there are also several measurements from patients whose hearts beat 3 times per second (180 bpm).

As it turns out, we can gain additional insight into what is going on by plotting not just the patient heart rate but also their ages when the measurements were recorded. Figure 3 uncovers the quirk: the dataset has newborn babies, and newborn babies have a markedly

different physiology from adults. Figure 3 is an interesting visualization of MIMIC, because it prompts discussion for three non-obvious artifacts:

1. Newborn physiology is significantly different from adult physiology.
2. Only one of MIMIC-III's two EHRs (CareVue) has data from a Neonatal ICU (NICU), whereas there are no newborns in the MetaVision data.
3. There are "300-year-old" patients in MIMIC. This is explained in the dataset's documentation: HIPAA requires all patients 90 years and older be binned together so as to avoid making patients easily identifiable by their rare age. MIMIC accomplishes this by mapping all 90+ patients to 300+.[3]

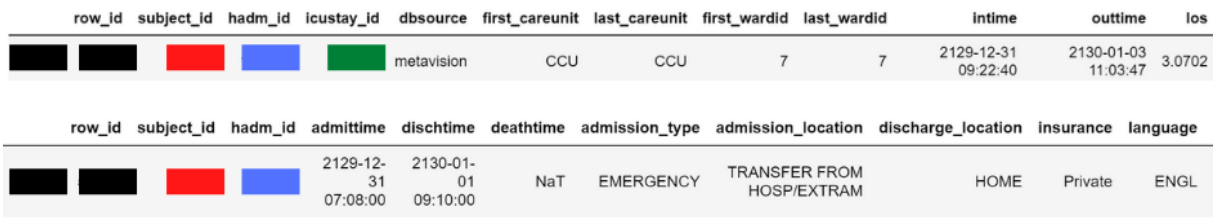
The presence of newborns and 300-year-old patients could potentially confuse someone not familiar with the patients in the dataset and how they were encoded. Many works that use MIMIC filter out patients younger than 15 years old (Ghassemi et al., 2017), however in a reproducibility study, MIMIC authors Johnson et al. observe that "[m]any of the studies reviewed omitted details necessary to fully reproduce the work [such as] which age to use for identifying adults" (Johnson et al., 2017).

A data scientist unfamiliar with what constitutes abnormal ranges for feature values might unwittingly just feed it into their model without further investigation. Ages of 0 and 300 will not crash the learning algorithm, though they likely may lead to the model learning incorrect generalizations and inferences.

Figure 3: Patient heart rate vs. patient age in the MIMIC-III Database. Patients 90+ are coded as being 300 years-old for de-identification purposes.

3. https://physionet.org/content/mimiciii/1.4

Figure 4: Data from the *icustays* (top) & *admissions* (bottom) in MIMIC-III of a patient who allegedly didn't leave the ICU until after being discharged from the hospital.



row_id	subject_id	hadm_id	icustay_id	dbsource	first_careunit	last_careunit	first_wardid	last_wardid	intime	outtime	los
[redacted]	[redacted]	[redacted]	[redacted]	metavision	CCU	CCU	7	7	2129-12-31 09:22:40	2130-01-03 11:03:47	3.0702

row_id	subject_id	hadm_id	admittime	dischtime	deathtime	admission_type	admission_location	discharge_location	insurance	language
[redacted]	[redacted]	[redacted]	2129-12-31 07:08:00	2130-01-01 09:10:00	NaT	EMERGENCY	TRANSFER FROM HOSP/EXTRAM	HOME	Private	ENGL

5.2. Inconsistent Timestamps

Figure 1 shows the data in the tables of MIMIC corresponds to a patient being treated in the ICU. When the patient arrives at the hospital, they are admitted for a hospital stay and their records are entered into the EHR. At some point during that stay the patient is sent to an intensive care unit (e.g., surgical ICU, trauma ICU, cardiac ICU, etc). They might have multiple ICU stays during one admission, and eventually they leave.

Intuitively, one would assume that a patient's admission in the hospital would follow the following chronological workflow:

hospital admission time (admittime) $\leq$
patient enters ICU (intime) $\leq$
patient leaves ICU (outtime) $\leq$
patient discharged from hospital (dischtime)

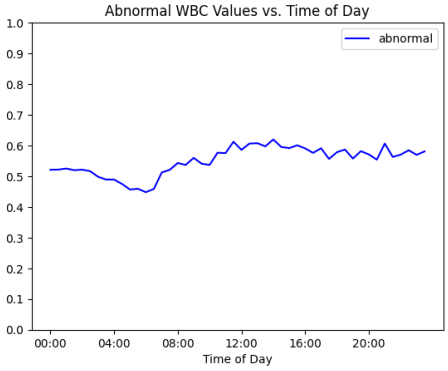
We examine 57,783 ICU stays and find that the above ordering (hosp-in, icu-in, icu-out, hosp-out) only strictly holds 79% of the time. There are 12,338 stays where at least one of those three inequalities does not hold:

- 486 instances of the ICU intime occuring before the hospital admission (i.e., admitted to ICU before admitted to hospital); and
- 11,886 instances of the ICU outtime occurring after the hospital dischtime (i.e., left the hospital before they left the ICU)

Our understanding is that this phenomenon is a reflection of data entry "errors." The hospital and ICU each have their own paperwork for when a patient leaves, and if it is filled out by different caregivers, then estimated times might be somewhat inconsistent with each other. For instance, nearly 3,000 of those 12,300 stays have the outtime and dischtime within 10 minutes of each other, suggesting this was the result of 2 sets of paperwork filled out by 2 different hospital workers both describing the same event. Further, 90% of those stays fall within 6 hours, which is half of a nurse's shift.

One of the first tasks in data science is to "clean" the data, where impossible data (e.g., a patient that is 20 feet tall) is filtered out. However, a nontrivial amount of patients (21%) do not fully satisfy the expected chronological relationships. "Common sense" cannot adjudicate whether those stays are acceptable or not; it requires understanding the generating process.

Figure 5: Fraction of abnormal results from white blood cell counts vs. time of day



5.3. Lab Values Vary by Time of Day

In 1816, Laennec invented the stethoscope to measure the human body and correlate those findings with the patient's disease. In the following centuries, medicine has developed thousands of new measurements for body temperature, reflexes, body chemistry, and more (Walker et al., 1990). Modern clinical laboratory tests are a product of that legacy: we try to measure the patient and deduce what that means about their health. According to a survey from the Los Angeles County/University of Southern California Medical Center, clinical laboratory tests are used for: diagnosis (37% of labs), monitoring (33%), screening (32%), comparing against previous abnormal result (12%), prognosis (7%), and misc (3%) (Wertman et al., 1980). The CDC reports 14 billion laboratory tests are ordered annually and "70% of today's medical decisions are depend on laboratory test results" (CDC, 2018).

However, it is important to note that the data captured in a patient's EHR is not their full story, but rather a collection of the data points chosen by the clinical care team and collected at specific times (Agniel et al., 2018). Albers and Hripcsak (2010) show that patients with kidney failure are more likely to have their creatinine measured between 10 pm and 6 am than healthier patients . Contrary to the naive assumption that data is missing randomly, the timing of the measurement can tell you a lot about the care team's assumptions and resources available.

In MIMIC, we observe that the fraction of normal vs. abnormal lab values are not always constant over time as one might naively expect. Figure 5 shows that lab abnormality[4] is not constant throughout the day. Between 4am and 8am, the majority of lab values are within normal range, but between 8am and 4am, there is a noticeably higher fraction of tests that are abnormal.

The implication of this finding is that the timing of a lab order might, itself, already be an indication of the patient's state. Some machine learning models are unequipped to handle irregularly-spaced sequential data, so time parameters are dropped for convenience. Of course, it is possible a given situation might permit a modeling decision like that, but in general, data scientists are not knowledgeable enough to understand those assessments without a domain expert.

4. where normal is defined as 4,500 - 11,000 white blood cells per microliter (UCSF, 2020)

Table 1: 1-year mortality rates by race in post-discharge cohort.

race	N	# 12-month mort	% 12-month mort
White	22482	3492	15.5
Not Specified	2860	420	14.7
Black	2435	370	15.2
Other	1825	171	9.4
Hispanic	1107	81	7.3
Asian	730	104	14.2
Total	31892	4699	14.7

5.4. Multiple Copies of Provider Notes

Throughout a patient's ICU stay, the care team writes many notes. Every day there are at least 2 nursing shifts, each with their own note. There is also typically a daily physician note. If certain tests are ordered, there might be radiology, echo, or ECG reports. If experts are called in, there might be nutrition or social worker notes. And finally, when they leave the hospital, there is a discharge summary.

Many papers employ the use of natural language processing (NLP) on hospital notes to predict clinical outcomes (Rumshisky et al., 2016). One approach clinical NLP predictive models take is concatenating all of the notes into a single note for processing (Boag et al., 2018a). However, if done naively, this runs the risk of double-, triple-, or even octuple-counting the content of these notes. This is because MIMIC contains not just a provider's official note but also all of the note drafts that the EHR autosaved. As a result, a clinical finding documented in one note might as if it is repeatedly observed. This manifests in the database in the following way: all of the partial drafts have the same "chart time" as the final note, but their "store times" reflect when they were partially written. This allows for a straightforward way to de-duplicate the set of notes for a patient: group by chart time and then select the most recently written note.

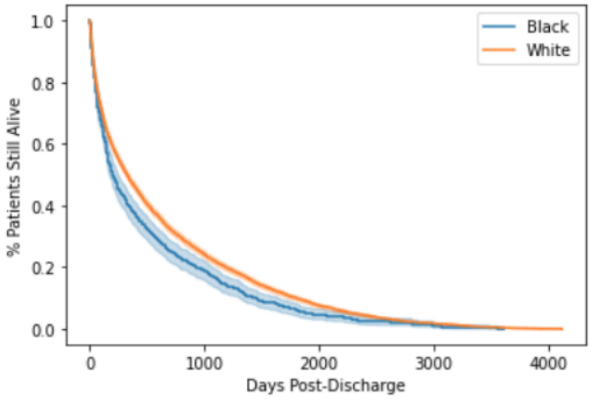
5.5. Missing Death Date Collection

Many end-of-life patients don't wish to spend their final weeks hooked up to machines in a hospital, but they often miss the opportunity to communicate how they'd like to live their final months (Gawande, 2014). Because "Unexpected end-of-life situations can happen at any age," the Mayo Clinic recommends that "it's important for all adults to prepare [advance directives]" (Staff', 2020). However, only 1-in-3 American adults have completed an advance directive (Yadav et al., 2017). To address this, there have been efforts to build machine learning models to predict which patients are at high risk and would benefit from discussing their goals of care (Avati et al., 2017).

We build a 12-month mortality model on MIMIC data to identify patients that would benefit most from advance care planning.[5] Our cohort contains adult patients who were discharged (i.e. didn't die in-hospital) and have never had a code status of "comfort mea-

5. https://github.com/wboag/eol-mort-pred

Figure 6: Kaplan Meier curve of patients that die postdischarge.



sures only" nor an ICD billing code for palliative care. Table 1 shows the demographic breakdown of this demographic. Surprisingly, in contrast with a large literature of racial biases in health (Williams, 1999), this data seems to show white patients have a higher 12-month mortality rate than Black patients.

Figure 6 shows a survival curve of post-discharge mortality and demonstrates that for any given time (e.g., 1-month post-discharge), a higher proportion of Black patients have died than white patients. This appears in direct conflict with Table 1 assertion that white patients have a higher 12-month mortality rate.

In order to understand this, we examine the data collection practices. MIMIC-III uses social security records to collect patient mortality information for patients who passed away outside of the hospital. But in 2013,[6] "a legislative change ... forbid the social security death index from collecting deaths from state databases. The result was a 40% drop in the capture of deaths - it's no longer of sufficient quality to use as a source of out-of-hospital mortality." Consequently, the date of death column is partially unreliable for patients post-discharge; a higher number of Black patients were passing away without that being reflected in the SSN's records (and therefore without being reflected in MIMIC).

This data label bias is easy to miss; we only identified it because of the surprising nature of the original cohort's mortality rates by race. However, training a model on the original cohort would produce a tool which recommends interventions that underestimate the risk for Black patients.[7] In contrast, training a model on a modified cohort (considering only patients who are known to have a death certificate eventually) is able to build a model which more accurately identifies that Black and Asian populations are at higher risk for 12-month mortality.[8] Intervention models that make directionally wrong mistakes run the risk of creating harmful feedback loops for patients.

6. https://github.com/MIT-LCP/mimic-code/issues/1199
7. https://github.com/wboag/eol-mort-pred/blob/main/eol-predict.ipynb
8. https://github.com/wboag/eol-mort-pred/blob/main/eol-predict-3.0.ipynb

6. Instances of Incorrect Assumptions

In this section, we examine two papers which made incorrect assumptions about MIMIC. Although neither of these assumptions are fatal to the papers' central claims, each error does undermine its paper's point when left unaddressed.

6.1. Diagnosis Column

In "What's In a Note?", Boag et al. (2018a) investigate different methods for representing clinical notes as high-dimensional vectors. Specifically, they employ three methods to represent the text of notes (bag-of-words, word2vec centroid, and LSTM) and explore how well each method does on predicting structured information like age, gender, diagnosis, and length of stay.

They found that bag-of-words (BoW) performed better on extractive tasks such as age and gender, where the answer was indicated by identifying a single, important token, such as in the sentence "Patient is a 38-year-old woman with fatigue and coughing." However, they also note that the "LSTM model outperforms BoW on tasks more related to clinical reasoning: diagnosis and length of stay, for which we expect the temporal information to be important in predictions." This work was yet another directionally consistent datapoint in the research about deep learning's value for difficult reasoning tasks (Hermann et al., 2015; Tai et al., 2015; See et al., 2017).

Although it may well be the case that the "LSTM model outperforms BoW on tasks more related to clinical reasoning ... which we expect the temporal information to be important in predictions," this paper is less supportive evidence than originally understood. The outcome they are predicting is not *final* diagnosis (which could be identified through ICD codes) but instead is the *admission* diagnosis, which is the condition identified by the physician when the patient first is admitted to the hospital. In this way, since the prediction label was assigned at the beginning of the hospital admission, there is no need for a model to "reason" through a sequence of data. Technically speaking, this oversight could have been foreseen with a thorough reading of the documentation[9] for the relevant table: "The DIAGNOSIS column provides a preliminary, free text diagnosis for the patient on hospital admission. The diagnosis is usually assigned by the admitting clinician and does not use a systematic ontology."

Fortunately, the impact of this oversight has been minimal. Of the 30 publicly available works which cite Boag et al., only five even mention the claim about LSTMs being better at clinical reasoning tasks (Jin et al., 2018; Kemp et al., 2019; Zhang et al., 2020a; Khan, 2019; Baruah, 2020). The vast majority of works cite this paper to either establish that researchers have been using notes for predictions or to comment about the relationship between structured and unstructured data. Two works cite the Boag et al. for its publicly available word2vec embeddings pre-trained on MIMIC (Zhang et al., 2020b; Flamholz et al., 2022).

9. https://mimic.mit.edu/docs/iii/tables/admissions

6.2. Variation in Data Entry

Whereas the previous example was an instance of not adequately reading the full documentation, there are also scenarios where researchers simply cannot know without working with frontline workers.

In the past decade, numerous studies have identified that nonwhite patients had worse end-of-life experiences than white patients (Muni et al., 2011; Lee et al., 2016; Hanchate et al., 2009). They looked at invasive care (e.g., mechanical ventilation) as compared to comfort-based care (i.e., hospice). White patients were receiving smaller amounts of invasive care than nonwhite – in particular Black and Hispanic – patients. Prior work hypothesizes that mistrust between the patient and the doctor could be one of the factors exacerbating the issue. Mistrust may lead a patient or family to question the intention of a care team's hospice recommendation (e.g., suspecting the hospital doesn't want to use resources), and instead demand additional invasive interventions (Garrett et al., 1993).

Researchers published a series of papers investigating whether this trust-based hypothesis could explain the difference in treatments. Both "Racial Disparities and Mistrust in End-of-Life Care" (Boag et al., 2018c) and "Modeling Mistrust in End-of-Life Care" (Boag et al., 2018b) attempted to model a proxy for the patient's trust in their care team. This was done by using coded interpersonal features from the chartevents table for all MIMIC patients, including: whether the patient was restrained, patient healthcare literacy, indication of family meetings, how thoroughly pain is being monitored, and more. They found that the consensus of three derived proxies for mistrust both: (1) seem to indicate an even larger high-trust vs low-trust treatment disparity than Black vs white, and (2) show that Black patients had higher levels of mistrust in their doctors.

However, in the years that followed the publication, we spoke to a nurse from Beth Israel about a different project, and we asked about how to understand the coded items in the chartevents table. She described how some of these entries come from an overwhelming number of pop-ups in the EHR which few people would have enough time to comprehensively fill out. Further, she said that data entry can vary from nurse to nurse and although some nurses do code family meetings in the chart, she has only ever indicated family meetings in her nursing notes. The implications of this revelation is that the prior work was potentially under-estimating the "mistrust score" proxy for some patients whose indications of tension were not coded in structured data. A more comprehensive modeling approach would use both the structured data and perform NLP to extract comparable concepts from the unstructured text.

Fortunately, for this very reason of epistemic humility/hedging, Boag et al. (2018c) purposefully avoided creating just one "definitive" proxy which could be misinterpreted as authoritative. Instead, they created three different scores for "mistrust," not all of which even used chartevents. Therefore this is little reason to doubt this data entry instance would change the overall findings of the work.

The original work was published with a medical doctor as a coauthor, although there were no nurses (let alone Beth Israel nurses) involved in the original project. A medical doctor likely would not have known the specific data entry culture for nurses. Once again, this serves as an important reminder to work with the frontline workers familiar with the data generating process.

7. Conclusion

In this work, we demonstrate many of the quirks of a well-known public healthcare dataset (MIMIC-III). We do not mean to single MIMIC-III out as uniquely unintuitive – if anything, its documentation is *much* better than most clinical datasets the authors are familiar with. Instead, we hope to convey that every dataset is only as meaningful as the researcher's knowledge of the data generating process. Whether it is incorrect timestamps, unexpected data entry by the frontline workers, or incomplete label collection, we want to underscore that data scientists can't do it alone (and shouldn't try to).

Indeed these challenges demonstrate a broader systems failure, because it is difficult for reviewers to assess whether a paper's assumptions seem reasonable. We applaud conferences like Machine Learning for Healthcare[10] ensuring every submitted paper is assessed by at least one technical reviewer and at least one clinical reviewer. Nonetheless even relevant reviewers would not be able to catch something like the data entry habits of a given hospital's nurses. For this reason, it becomes even more critical to collaborate with a domain expert to help paint a clear picture of the data collection process.

Acknowledgments

First and foremost, thank you to the MIT Laboratory for Computational Physiology (LCP) and PhysioNet for creating and maintaining MIMIC, which has been a terrific resource for educators and researchers. In particular, thank you to Alistair Johnson for so much work creating MIMIC and carefully documenting it (even if we and other sometimes forget to read the docs). Additionally, thank you to the MIT Clinical Decision-Making Group (MEDG) for their iterative feedback on this work as it evolved.

References

Denis Agniel, Isaac Kohane, and Griffin Weber. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ*, 2018. doi: 10.1136/bmj.k4416. URL https://www.bmj.com/content/bmj/361/bmj.k1479.full.pdf.

D.j. Albers and George Hripcsak. A statistical dynamics approach to the study of human health data: Resolving population scale diurnal variation in laboratory data. *Physics Letters A*, 374(9):1159–1164, 2010. doi: 10.1016/j.physleta.2009.12.067.

Anand Avati, Kenneth Jung, Stephanie Harman, Lance Downing, Andrew Ng, and Nigam H. Shah. Improving palliative care with deep learning. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 311–316, 2017. doi: 10.1109/BIBM.2017.8217669.

Prakrit Baruah. *Predicting Hospital Readmission using Unstructured Clinical Note Data*. Master's thesis, Brown University, Providence, RI, 2020.

10. https://www.mlforhc.org/paper-submission

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives, 2014.

Willie Boag, Dustin Doss, Tristan Naumann, and Peter Szolovits. What's in a note? unpacking predictive value in clinical note representations. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2017:26–34, 05 2018a.

Willie Boag, Harini Suresh, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Modeling mistrust in end-of-life care. *Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2018)*, 2018b.

Willie Boag, Harini Suresh, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Racial disparities and mistrust in end-of-life care. In Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 3rd Machine Learning for Healthcare Conference*, volume 85 of *Proceedings of Machine Learning Research*, pages 587–602. PMLR, 17–18 Aug 2018c. URL https://proceedings.mlr.press/v85/boag18a.html.

Meredith Broussard. *Artificial unintelligence: How Computers Misunderstand the World*. The MIT Press, 2019.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA, 02 2018. PMLR. URL http://proceedings.mlr.press/v81/buolamwini18a.html.

CDC. Strengthening clinical laboratories, Nov 2018. URL https://www.cdc.gov/csels/dls/strengthening-clinical-labs.html.

C. D'Ignazio and L.F. Klein. *Data Feminism*. Strong Ideas. MIT Press, 2020. ISBN 9780262044004. URL https://books.google.com/books?id=x5nSDwAAQBAJ.

Zachary N. Flamholz, Andrew Crane-Droesch, Lyle H. Ungar, and Gary E. Weissman. Word embeddings trained on published case reports are lightweight, effective for clinical tasks, and free of protected health information. *Journal of Biomedical Informatics*, 125: 103971, 2022. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2021.103971. URL https://www.sciencedirect.com/science/article/pii/S1532046421003002.

Joanne Mills Garrett, Russell P. Harris, Jean K. Norburn, Donald L. Patrick, and Marion Danis. Life-sustaining treatments during terminal illness - who wants what? *Journal of General Internal Medicine*, 8(7):361–368, 7 1993. ISSN 0884-8734. doi: 10.1007/BF02600073.

Atul Gawande. *Being Mortal: Medicine and What Matters in the End*. Picador, 2014.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *CoRR*, abs/1803.09010, 2018. URL http://arxiv.org/abs/1803.09010.

Marzyeh Ghassemi, Mike Wu, Michael C. Hughes, Peter Szolovits, and Finale Doshi-Velez. Predicting intervention onset in the icu with switching state space models. *AMIA Summits on Translational Science Proceedings*, 2017:82 – 91, 2017.

Amresh Hanchate, Andrea C. Kronman, Yinong Young-Xu, Arlene S. Ash, and Ezekiel Emanuel. Racial and ethnic differences in end-of-life costs: Why do minorities cost more than whites? *Archives of Internal Medicine*, 169(5):493–501, 2009.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1), Jun 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0103-9. URL http://dx.doi.org/10.1038/s41597-019-0103-9.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1693–1701, Cambridge, MA, USA, 2015. MIT Press.

Mengqi Jin, Mohammad Taha Bahadori, Aaron Colak, Parminder Bhatia, Busra Celikkaya, Ram Bhakta, Selvan Senthivel, Mohammad Khalilia, Daniel Navarro, Borui Zhang, Tiberiu Doman, Arun Ravi, Matthieu Liger, and Taha A. Kass-Hout. Improving hospital mortality prediction with medical named entities and multimodal learning. *ArXiv*, abs/1811.12276, 2018.

Alistair E 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.

Alistair E. W. Johnson, Tom J. Pollard, and Roger G. Mark. Reproducibility in critical care: a mortality prediction case study. In Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 361–376. PMLR, 18–19 Aug 2017. URL https://proceedings.mlr.press/v68/johnson17a.html.

Jonas Kemp, Alvin Rajkomar, and Andrew M. Dai. Improved patient classification with language model pretraining over clinical notes. *CoRR*, abs/1909.03039, 2019. URL http://arxiv.org/abs/1909.03039.

Mohammad Hashir Khan. *A CNN-LSTM for predicting mortality in the ICU*. Master's thesis, University of Tennessee, Knoxville, Knoxville, TN, 2019.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Quoc V. Le, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg S. Corrado, Jeff Dean, and Andrew Y. Ng. Building high-level features using large scale unsupervised learning, 2012.

Janet J. Lee, Ann C. Long, J. Randall Curtis, and Ruth A. Engelberg. The influence of race/ethnicity and education on family ratings of the quality of dying in the icu. *Journal of Pain and Symptom Management*, 51(1):9 – 16, 2016. ISSN 0885-3924. doi: https://doi.org/10.1016/j.jpainsymman.2015.08.008. URL http://www.sciencedirect.com/science/article/pii/S0885392415004558.

Sarah Muni, Ruth A. Engelberg, Patsy D. Treece, Danae Dotolo, and J. Randall Curtis. The influence of race/ethnicity and socioeconomic status on end-of-life care in the icu. *Chest*, 139(5):1025–1033, 2011. ISSN 0012-3692. doi: https://doi.org/10.1378/chest.10-3011. URL https://www.sciencedirect.com/science/article/pii/S0012369211602262.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366: 447–453, 10 2019. doi: 10.1126/science.aax2342.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, page 79–86, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1118693.1118704. URL https://doi.org/10.3115/1118693.1118704.

Ravi B. Parikh, Christopher Manz, Corey Chivers, Susan Harkness Regli, Jennifer Braun, Michael E. Draugelis, Lynn M. Schuchter, Lawrence N. Shulman, Amol S. Navathe, Mitesh S. Patel, and Nina R. O'Connor. Machine Learning Approaches to Predict 6-Month Mortality Among Patients With Cancer. *JAMA Network Open*, 2(10):e1915997–e1915997, 10 2019. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2019.15997. URL https://doi.org/10.1001/jamanetworkopen.2019.15997.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

A Rumshisky, M Ghassemi, T Naumann, P Szolovits, V M Castro, T H Mccoy, and R H Perlis. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Translational Psychiatry*, 6(10), 2016. doi: 10.1038/tp.2015.182.

Shems Saleh, William Boag, Lauren Erdman, and Tristan Naumann. Clinical collabsheets: 53 questions to guide a clinical collaboration. In Finale Doshi-Velez, Jim Fackler, Ken

Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 5th Machine Learning for Healthcare Conference*, volume 126 of *Proceedings of Machine Learning Research*, pages 783–812. PMLR, 07–08 Aug 2020. URL https://proceedings.mlr.press/v126/saleh20a.html.

Abigail See, Peter Liu, and Christopher Manning. Get to the point: Summarization with pointer-generator networks. In *Association for Computational Linguistics*, 2017. URL https://arxiv.org/abs/1704.04368.

Jennifer L. Skeem and Christopher Lowenkamp. Risk, race, & recidivism: Predictive bias and disparate impact, June 2016. URL https://ssrn.com/abstract=2687339.

"Mayo Clinic Staff". Living wills and advance directives for medical decisions. Mayo Foundation for Medical Education and Research, 2020. URL https://www.mayoclinic.org/healthy-lifestyle/consumer-health/in-depth/living-wills/art-20046303.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1150. URL https://aclanthology.org/P15-1150.

John Tukey. *Exploratory Data Analysis*. Pearson, 1977.

UCSF. Wbc count, Oct 2020. URL https://www.ucsfhealth.org/medical-tests/wbc-count#:~:text=ThenormalnumberofWBCs,ormaytestdifferentspecimens.

H Kenneth Walker, W Dallas Hall, and J Willis Hurst. *Clinical Methods: The History, Physical, and Laboratory Examinations*. Butterworths, 1990. ISBN 0-409-90077-X. URL https://www.ncbi.nlm.nih.gov/books/NBK201/.

Shirly Wang, Matthew B. A. Mcdermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C. Hughes, and Tristan Naumann. Mimic-extract. *Proceedings of the ACM Conference on Health, Inference, and Learning*, 2020. doi: 10.1145/3368555.3384469. URL https://www.michaelchughes.com/papers/WangMcDermottEtAl_CHIL_2020.pdf.

Bradley G. Wertman, Stuart V. Sostrin, Zdena Pavlova, and George D. Lundberg. Why Do Physicians Order Laboratory Tests?: A Study of Laboratory Test Request and Use Patterns. *JAMA*, 243(20):2080–2082, 05 1980. ISSN 0098-7484. doi: 10.1001/jama.1980.03300460054033. URL https://doi.org/10.1001/jama.1980.03300460054033.

David R. Williams. Race, socioeconomic status, and health the added effects of racism and discrimination. *Annals of the New York Academy of Sciences*, 896(1):173–188, 1999. ISSN 1749-6632. doi: 10.1111/j.1749-6632.1999.tb08114.x. URL http://dx.doi.org/10.1111/j.1749-6632.1999.tb08114.x.

Kuldeep N. Yadav, Nicole B. Gabler, Elizabeth Cooney, Saida Kent, Jennifer Kim, Nicole Herbst, Adjoa Mante, Scott D. Halpern, and Katherine R. Courtright. Approximately

one in three us adults completes any type of advance directive for end-of-life care. *Health Affairs*, 36(7):1244–1251, 2017. doi: 10.1377/hlthaff.2017.0175. URL https://doi.org/10.1377/hlthaff.2017.0175.

Eddie Yang and Margaret E. Roberts. Censorship of online encyclopedias: Implications for nlp models. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 537–548, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445916. URL https://doi.org/10.1145/3442188.3445916.

John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15 (11):e1002683, nov 2018. doi: 10.1371/journal.pmed.1002683. URL https://doi.org/10.1371%2Fjournal.pmed.1002683.

Dongyu Zhang, Jidapa Thadajarassiri, Cansu Sen, and Elke Rundensteiner. Time-aware transformer-based network for clinical notes series prediction. In Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 5th Machine Learning for Healthcare Conference*, volume 126 of *Proceedings of Machine Learning Research*, pages 566–588. PMLR, 07–08 Aug 2020a. URL https://proceedings.mlr.press/v126/zhang20c.html.

Haoran Zhang, Elisa Candido, Andrew S. Wilton, Raquel Duchen, Liisa Jaakkimainen, Walter Wodchis, and Quaid Morris. *Identifying Transitional High Cost Users from Unstructured Patient Profiles Written by Primary Care Physicians*, pages 127–138. World Scientific, 2020b. doi: 10.1142/9789811215636_0012. URL https://www.worldscientific.com/doi/abs/10.1142/9789811215636_0012.