Textual Evidence for the Perfunctoriness of Independent Medical Reviews

Adrian Brasoveanu
abrsvn@ucsc.edu
University of California Santa Cruz
Santa Cruz, CA

Megan Moodie
mmoodie@ucsc.edu
University of California Santa Cruz
Santa Cruz, CA

Rakshit Agrawal
ragrawal@camio.com
Camio Inc.
San Mateo, CA

ABSTRACT

We examine a database of 26,361 Independent Medical Reviews (IMRs) for privately insured patients, handled by the California Department of Managed Health Care (DMHC) through a private contractor. IMR processes are meant to provide protection for patients whose doctors prescribe treatments that are denied by their health insurance (either private insurance or the insurance that is part of their worker comp; we focus on private insurance here). Laws requiring IMR were established in California and other states because patients and their doctors were concerned that health insurance plans deny coverage for medically necessary services. We analyze the text of the reviews and compare them closely with a sample of 50000 Yelp reviews [19] and the corpus of 50000 IMDB movie reviews [10]. Despite the fact that the IMDB corpus is twice as large as the IMR corpus, and the Yelp sample contains almost twice as many reviews, we can construct a very good language model for the IMR corpus using inductive sequential transfer learning, specifically ULMFiT [8], as measured by the quality of text generation, as well as low perplexity (11.86) and high categorical accuracy (0.53) on unseen test data, compared to the larger Yelp and IMDB corpora (perplexity: 40.3 and 37, respectively; accuracy: 0.29 and 0.39). We see similar trends in topic models [17] and classification models predicting binary IMR outcomes and binarized sentiment for Yelp and IMDB reviews. We also examine four other corpora (drug reviews [6], data science job postings [9], legal case summaries [5] and cooking recipes [11]) to show that the IMR results are not typical for specialized-register corpora. These results indicate that movie and restaurant reviews exhibit a much larger variety, more contentful discussion, and greater attention to detail compared to IMR reviews, which points to the possibility that a crucial consumer protection mandated by law fails a sizeable class of highly vulnerable patients.

CCS CONCEPTS

• **Computing methodologies** → **Latent Dirichlet allocation**; **Neural networks**.

KDD-KiML'20, August 2020, San Diego, CA, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

KEYWORDS

AI for social good, state-managed medical review processes, language models, topic models, sentiment classification

ACM Reference Format:
Adrian Brasoveanu, Megan Moodie, and Rakshit Agrawal. 2020. Textual Evidence for the Perfunctoriness of Independent Medical Reviews. In *Proceedings of KDD-KiML (KDD-KiML'20).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

1.1 Origin and structure of IMRs

Independent Medical Review (IMR) processes are meant to provide protection for patients whose doctors prescribe treatments that are denied by their health insurance – either private insurance or the insurance that is part of their workers' compensation. In this paper, we focus exclusively on privately insured patients. Laws requiring IMR processes were established in California and other states in the late 1990s because patients and their doctors were concerned that health insurance plans deny coverage for medically necessary services to maximize profit.[1]

As aptly summarized in [1], IMR is regularly used to settle disputes between patients and their health insurers over what is medically necessary or experimental/investigational care. Medical necessity disputes occur between health plans and patients because the health plan disagrees with the patient's doctor about the appropriate standard of care or course of treatment for a specific condition. Under the current system of managed care in the U.S., services rendered by a health care provider are reviewed to determine whether the services are medically necessary, a process referred to as utilization review (UR). UR is the oversight mechanism through which private insurers control costs by ensuring that only medically necessary care, covered under the contractual terms of a patient's insurance plan, is provided. Services that are not deemed medically necessary or fall outside a particular plan are not covered.

Procedures or treatment protocols are deemed experimental or investigational because the health plan – but not necessarily the patient's doctor, who in many cases has enough clinical confidence in a treatment to order it – considers them non-routine medical care, or takes them to be scientifically unproven to treat the specific condition, illness, or diagnosis for which their use is proposed.

[1] For California, see the Friedman-Knowles Act of 1996, requiring California health plans to provide external independent medical review (IMR) for coverage denials. As of late 2002, 41 states and the District of Columbia had passed legislation creating an IMR process. In 34 of these states, including California, the decision resulting from the IMR is binding to the health plan. See [1, 15] for summaries of the political and legal history of the IMR system, and [2] for an early partial survey of the DMHC IMR data.

It is important to realize that the IMR process is usually the third and final stage in the medical review process. The typical progression is as follows. After in-person and possibly repeated examination of the patient, the doctor recommends a treatment, which is then submitted for approval to the patient's health plan. If the treatment is denied in this first stage, both the doctor and the patient may file an appeal with the health plan, which triggers a second stage of reviews by the health-insurance provider, for which a patient can supply additional information and a doctor may engage in what is known as a "peer to peer" discussion with a health-insurance representative. If these second reviews uphold the initial denial, the only recourse the patient has is the state-regulated IMR process, and per California law, an IMR grievance form (and some additional information) is included with the denial letter.

An IMR review must be initiated by the patient and submitted to the California Department of Managed Health Care (DMHC), which manages IMRs for privately-insured patients. Motivated treating physicians may provide statements of support for inclusion in the documentation provided to DMHC by the patient, but in theory the IMR creates a new relationship of care between the reviewing physician(s) hired by a private contractor on behalf of DMHC, and the patient in question. The reviewing physicians' decision is supposed to be made based on what is in the best interest of the patient, not on cost concerns. It is this relation of care that constitutes the consumer protection for which IMR processes were legislated. Understandably, given that the patients in question may be ill or disabled or simply discouraged by several layers of cumbersome bureaucratic processes, there is a very high attrition from the initial review to the final, IMR, stage. That is, only the few highly motivated and knowledgeable patients – or the extremely desperate – get as far as the IMR process.

The IMR process is regulated by the state, but it is actually conducted by a third party. At this time (2019), the provider in California and several other states across the US is MAXIMUS Federal Services, Inc.[2] The costs associated with the IMR review, at least in California, are covered by health insurers. It is DMHC's and MAXIMUS's responsibility to collect all the documentation from the patient, the patient's doctor(s) and the health insurer. There are no independent checks that all the documentation has actually been collected, however, and patients do not see a final list of what has been provided to the reviewer prior to the IMR decision itself (a *post facto* list of file contents is mailed to patients along with the final, binding, decision; it is unclear what recourse a patient may have if they find pertinent information was missing from the review file). Once the documentation is assembled, MAXIMUS forwards it to anywhere from one to three reviewers, who remain anonymous, but are certified by MAXIMUS to be appropriately credentialed and knowledgeable about the treatment(s) and condition(s) under review. The reviewer submits a summary of the case, and also a rationale and evidence in support of their decision, which is a binary *Upheld/Overturned* decision about the medical service. IMR reviewers do not enter a consultative relationship with the patient, doctor or health plan – they must render an uphold/overturn decision based solely on the provided medical records. However, as noted above, they are in an implied relationship of care to the patient, a point to which we return in the Discussion section below (§4).

While insurance carriers do not provide statistics about the percentage of requested treatments that are denied in the initial stage, looking at the process as a whole, a pattern of service denial aimed to maximize profit, rather than simply maintain cost effectiveness, seems to emerge. Typically, the argument for denial contends that the evidence for the beneficial effects of the treatment fails the prevailing standard of scientific evidence. This prevailing standard invoked by IMR reviewers is usually randomized control trials (RCTs), which are expensive, time-consuming trials that are run by large pharmaceutical companies only if the treatment is ultimately estimated to be profitable.

RCTs, however, have known limits: they "require minimal assumptions and can operate with little prior knowledge [which] is an advantage when persuading distrustful audiences, but it is a disadvantage for cumulative scientific progress, where prior knowledge should be built upon, not discarded." [3] Inflexibly applying the RCT "gold standard" in the IMR process is often a way to ignore the doctors' knowledge and experience in a way that seems superficially well-reasoned and scientific. "RCTs can play a role in building scientific knowledge and useful predictions" – and we add, treatment recommendations – "only [...] as part of a cumulative program, [in combination] with other methods." [3]

Notably, the experimental/investigational category of treatments that get denied often includes promising treatments that have not been fully tested in clinical RCTs – because the treatment is new or the condition is rare in the population, so treatment development costs might not ultimately be recovered. Another common category of experimental/investigational denials involves "off-label" drug uses, that is, uses of FDA-approved pharmaceuticals for a purpose other than the narrow one for which the drug was approved.

1.2 Main argument and predictions

Recall that these *'experimental' treatments or off-label uses are recommended by the patient's doctor*, and therefore their potential benefits are taken to outweigh their possible negative effects. The recommending doctor is likely very familiar with the often lengthy, tortuous and *highly specific medical history of the patient*, and with the list of 'less experimental' treatments that have been proven unsuccessful or have been removed from consideration for patient-specific reasons. It is also important to remember that many *rare conditions have no "on-label" treatment options available*, since expensive RCTs and treatment approval processes are not undertaken if companies do not expect to recover their costs, which is likely if the potential 'market' is small (few people have the rare condition).

Therefore, our main line of argumentation is as follows.

- Since IMRs are the final stage in a long bureaucratic process in which health insurance companies keep denying coverage for a treatment repeatedly recommended by a doctor as medically necessary, we expect that the issue of medical necessity is non-trivial when that specific patient and that specific treatment are carefully considered.
- We should therefore expect the text of the IMRs, which justifies the final determination, to be highly individualized and argue for that final decision (whether congruent with the

[2] https://www.maximus.com/capability/appeals-imr

health plan's decision or not) in a way that involves the particulars of the treatment and the particulars of the patient's medical history and conditions.

Thus, we expect a reasoned, thoughtful IMR to *not be highly generic and templatic / predictable* in nature. For instance, legal documents may be highly templatic as they discuss the application of the same law or policy across many different cases, but a response carefully considering the specifics of a medical case reaching the IMR stage is not likely to be similar to many other cases. We only expect high similarity and 'templaticity' for IMR reviews if they are reduced to a more or less automatic application of some prespecified set of rules (rubber-stamping).

1.3 Main results, and their limits

Concomitantly with this quantitative study, we conducted preliminary qualitative research with a focus on pain management and chronic conditions. We investigated the history of the IMR process, in addition to having direct experience with it. We had detailed conversations with doctors in Northern California and on private social media groups formed around chronic conditions and pain management. This preliminary research reliably points towards the possibility that IMR reviews are perfunctory, and that this crucial consumer protection mandated by law seems to fail for a sizeable class of highly vulnerable patients. In this paper, we focus on the text of the IMR decisions and attempt to quantify the evidence for the perfunctoriness of the IMR process that they provide.

The text of the IMR findings does not provide unambiguous evidence about the quality and appropriateness of the IMR process. If we had access to the full, anonymized patient files submitted to the IMR reviewers (in addition to the final IMR decision and the associated text), we might have been able to provide much stronger evidence that IMRs should have a significantly higher percentage of overturns, and that the IMR process should be improved in various ways, e.g., (*i*) patients should be able to check that all the relevant documentation has been collected and will be reviewed, and (*ii*) the anonymous reviewers should be held to higher standards of doctor-patient care. At the very least, one would want to compare the reports/letters produced by the patient's doctor(s) and the IMR texts. However, such information is not available and there are no visible signs suggesting potential availability in the near future. The information that is made available by DMHC constitutes the IMR decision – whether to uphold or overturn the health plan decision –, the anonymized decision letter, and information about the requested treatment category (also available in the letter). We, therefore, had to limit ourselves to the text of the DMHC-provided IMR findings in our empirical analysis.

A qualitative inspection of the corpus of IMR decisions made available by the California DMHC site as of June 2019 (a total of 26631 cases spanning the years 2001-2019) indicates that the reviews – as documented in the text of the findings – focus more on the review procedure and associated legalese than on the actual medical history of the patient and the details of the case. For example, decisions for chronic pain management seem to mostly rubber-stamp the Medical Treatment Utilization Schedule (MTUS) guidelines, with very little consideration of the rarity of the underlying condition(s) (see our comments about RCTs above), or a thoughtful evaluation of the risk/benefit profile of the denied treatment relative to the specific medical history of the patient (assuming this history was adequately documented to begin with).

The goal in this paper is to investigate to what extent Natural Language Processing (NLP) / Machine Learning (ML) methods that are able to extract insights from large corpora point in the same direction, thus mitigating cherry-picking biases that are sometimes associated with qualitative investigations. In addition to the IMR text, we perform a comparative study with additional English-language datasets in an attempt to eliminate data-specific and problem-specific biases.

- We analyze the text of the IMR reviews and compare them with a sample of 50,000 Yelp reviews [19] and the corpus of 50,000 IMDB movie reviews [10].
- As the size of data has significant consequences for language-model training, and NLP/ML models more generally, we expect models trained on the Yelp and IMDB corpora to outperform models trained on the IMR corpus, given that the IMDB corpus is twice as large as the IMR corpus, and the Yelp samples contain almost twice as many reviews.
- In this paper, we instead demonstrate that we were able to construct a very good language model for the IMR corpus using inductive sequential transfer learning, specifically ULMFiT [8], as measured by the quality of text generation.
- In addition, the model achieves a much lower perplexity (11.86) and a higher categorical accuracy (0.53) on unseen test data, compared to models trained on the larger Yelp and IMDB corpora (perplexity: 40.3 and 37, respectively; categorical accuracy: 0.29 and 0.39).
- We see similar trends in topic models [17] and classification models predicting binary IMR outcomes and binarized sentiment for Yelp and IMDB reviews.

These results indicate that movie and restaurant reviews exhibit a much larger variety, more contentful discussion, and greater attention to detail compared to IMR reviews. In an attempt to mitigate confirmation bias, as well as potentially significant register differences between IMRs and movie or restaurant reviews, we examine four additional corpora: drug reviews [6], data science job postings [9], legal case summaries [5] and cooking recipes [11]. These specialized-register corpora are potentially more similar to IMRs than IMDB or Yelp: the texts are more likely to be highly similar, include boilerplate text and have a templatic/standardized structure. We find that predictability of IMR texts, as measured by language-model perplexity and categorical accuracy, is higher than all the comparison datasets by a good margin.

Based on these empirical comparisons, we conclude that we have strong evidence that the IMR reviews are perfunctory and, therefore, that a crucial consumer protection mandated by law seems to fail for a sizeable class of highly vulnerable patients. The paper is structured as follows. In Section 2, we discuss the datasets in detail, with a focus on the nature and characteristics of the IMR data. In Section 3, we discuss the models we use to analyze the IMR, Yelp and IMDB datasets, as well as the four auxiliary corpora (drug reviews, data science jobs, legals cases and recipes). The section also compares and discusses the results of these models. Section 4 puts all the results together into an argument for the perfunctoriness of

the IMRs. Section 5 concludes the paper and outlines directions for future work.

2 THE DATASETS

2.1 The IMR dataset

The IMR dataset was obtained from the DMHC website in June 2019[3] and was minimally preprocessed. It contains 26,361 cases / observations and 14 variables, 4 of which are the most relevant:

- TreatmentCategory: the main treatment category;
- ReportYear: year the case was reported;
- Determination: indicates if the determination was upheld or overturned;
- Findings: a summary of the case findings.

The top 14 treatment categories (with percentages of total ≥ 2%), together with their raw counts and percentages are provided in Table 1.

Table 1: Top 14 treatment categories

TreatmentCategory	Case count	% of total
Pharmacy	6480	25%
Diag Imag & Screen	4187	16%
Mental Health	2599	10%
DME	1714	7%
Gen Surg Proc	1227	5%
Orthopedic Proc	1173	5%
Rehab/ Svc - Outpt	1157	4%
Cancer Care	1029	4%
Elect/Therm/Radfreq	828	3%
Reconstr/Plast Proc	825	3%
Autism Related Tx	767	3%
Emergency/Urg Care	582	2%
Diag/ MD Eval	573	2%
Pain Management	527	2%

The breakdown of cases by patient gender (not recorded for all cases) is as follows: Female – 14823 (56%), Male – 10836 (41%), Other – 11 (0.0004%).

The breakdown by determination (the outcome of the IMR) is: Upheld – 14309 (54%), Overturned – 12052 (46%).

The outcome counts and percentages by year are provided in Table 2. The number of cases for 2019 include only the first 5 months of the year plus a subset of June 2019.

Interestingly, the DMHC website featured a graphic in June 2019 (Figure 1) that reports the percentage of Overturned outcomes to be 64%, a figure that does not accord with any of our data summaries. We intend to follow up on this issue and see if the DMHC can share their data-analysis pipeline so that we can pinpoint the source(s) of this difference.

Given that our main goal here is to investigate the text of the IMR findings and its predictiveness with respect to IMR outcomes, we provide some general properties of this corpus. The histogram of word counts for the IMR findings (the text associated with each case) is provided in Figure 2. There are 26,361 texts, with a total of 5,584,280 words. Words are identified by splitting texts on white space (sufficient for our purposes here). The mean length of a text is 211.84 words, with a standard deviation (SD) of 120.58.

[3] https://data.chhs.ca.gov/dataset/independent-medical-review-imr-determinations-trend.

Table 2: Outcome counts and percentages by year

ReportYear	Total # of cases	Overturned	Upheld
2001	28	7 (25%)	21
2002	695	243 (35%)	452
2003	738	280 (38%)	458
2004	788	305 (39%)	483
2005	959	313 (33%)	646
2006	1080	442 (41%)	638
2007	1342	571 (43%)	771
2008	1521	678 (45%)	843
2009	1432	641 (45%)	791
2010	1453	661 (45%)	792
2011	1435	684 (48%)	751
2012	1203	589 (49%)	614
2013	1197	487 (41%)	710
2014	1433	549 (38%)	884
2015	2079	1070 (51%)	1009
2016	3055	1714 (56%)	1341
2017	2953	1391 (47%)	1562
2018	2545	1218 (48%)	1327
2019	425	209 (49%)	216

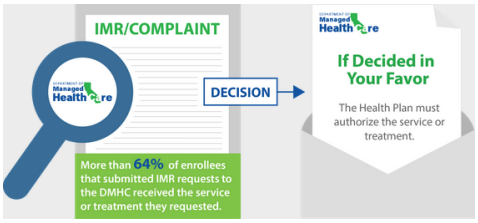


Figure 1: % Overturned claimed on DMHC site (June 2019)

2.2 The comparison datasets

As comparison datasets, we use the IMDB movie-review dataset [10], which has 50,000 reviews and a binary positive/negative sentiment classification associated with each review. This dataset will be particularly useful as a baseline for our ULMFiT transfer-learning language models (and subsequent transfer-learning classification models), where we show that we obtain results for the IMDB dataset that are similar to the ones in the original ULMFiT paper [8].

There are 50,000 movie reviews in the IMDB dataset, evenly split into negative and positive reviews. The histogram of text lengths for IMDB reviews is provided in Figure 2. The reviews contain a total of 11,557,297 words. The mean length of a review is 231.15 words, with an SD of 171.32.

We select a sample of 50,000 Yelp (mainly restaurant) reviews [19], with associated binarized negative/positive evaluations, to provide a comparison corpus intermediate between our DMHC dataset and the IMDB dataset. From a total of 560,000 reviews (evenly split between negative and positive), we draw a weighted random sample with the weights provided by the histogram of text lengths for the IMR corpus. The resulting sample contains 25,809 (52%) negative reviews and 24,191 (48%) positive reviews. The histogram of text

lengths for Yelp reviews is also provided in Figure 2. The reviews contain a total of 7,038,467 words. The mean length of a review is 140.77 words, with an SD of 71.09.

2.3 Four auxiliary datasets

We will also analyze four other specialized-register corpora: drug reviews [6], data science (DS) job postings [9], legal case reports [5] and cooking recipes [11]. The modeling results for these specialized-register corpora will enable us to better contextualize and evaluate the modeling results for the IMR, IMDB and Yelp corpora, since these four auxiliary datasets might be seen as more similar to the IMR corpus than movie or restaurant reviews. The drug-review corpus contains reviews of pharmaceutical products, which are closer in subject matter to IMRs than movie/restaurant reviews. The other three corpora are all highly specialized in register, just like the IMRs, with two of them (DS jobs and legal cases) particularly similar to the IMRs in that they involve templatic texts containing information aimed at a specific professional sub-community.

These four corpora are very different from each other and from the IMR corpus in terms of (*i*) the number of texts that they contain and (*ii*) the average text length (number of words per text). Because of this, there was no obvious way to sample from them and from the IMR, IMDB and Yelp corpora in such a way that the resulting samples were both roughly comparable with respect to the total number of texts and average text length, and also large enough to obtain reliable model estimates. We therefore analyzed these four corpora as a whole.

The drug-review corpus includes 132,300 drugs reviews – more than the double the number of texts in the IMDB and Yelp datasets, and more than 4 times the number of texts in the IMR dataset. From the original corpus of 215,063 reviews, we only retained the reviews associated with a rating of 10, which we label as *positive* reviews, and a rating of 1 through 5, which we label as *negative* reviews.[4]

The histogram of text lengths for drug reviews is provided in Figure 3. The reviews contain a total of 11,015,248 words, with a mean length of 83.26 words per review (significantly shorter than the IMR/IMDB/Yelp texts) and an SD of 45.73.

The DS corpus includes 6,953 job postings (about a quarter of the texts in the IMR corpus), with a total of 3,731,051 words. The histogram of text lengths is provided in Figure 3. The mean length of a job posting is 536.61 words (more than twice as long as the IMR/IMDB/Yelp texts), with an SD of 254.06.

There are 3,890 legal-case reports (even fewer than DS job postings), with a total of 25,954,650 words (about 5 times larger than the IMR corpus). The histogram of text lengths for the legal-case reports is provided in Figure 3. The mean length of a report is 6,672.15 words (a degree of magnitude longer than IMR/IMDB/Yelp), with a very high SD of 11,997.98.

Finally, the recipe corpus includes more than 1 million texts: there are 1,029,719 recipes, with a total of 117,563,275 words (very large compared to our other corpora). The histogram of text lengths for the recipes is provided in Figure 3. The mean length of a recipe is 114.17 words (close to the length of a drug review, and roughly half of an IMR), with an SD of 90.54.

[4]We did this so that we have a fairly balanced dataset (68,005 *positive* drug reviews and 64,295 *negative* reviews) to estimate classification models like the ones we report for the IMR, IMDB and Yelp corpora in the next section. For completeness, the drug-review classification results on previously unseen test data are as follows: logistic regression accuracy: 77.89%; accuracy of multilayer perceptron with a 1,000-unit hidden layer and a ReLU non-linearity: 83.18%; ULMFiT classification model accuracy: 96.12%.

3 THE MODELS

In this section, we analyze the text of the IMR findings and its predictiveness with respect to IMR outcomes. We systematically compare these results with the corresponding ones for the IMDB and Yelp corpora. The datasets were split into training (80%), validation (10%) and test (10%) sets. Test sets were only used for the final model evaluation.

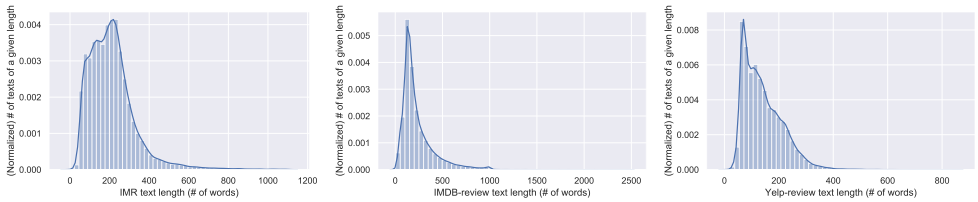
We start with baseline classification models (logistic regressions and logistic multilayer perceptrons with one hidden layer) to establish that the reviews in all three datasets under consideration are highly predictive of the associated binary outcomes. Once the predictiveness, hence, relevance, of the text is established, we turn to an in-depth analysis of the texts themselves by means of topic and language models. We see that the text of the IMR reviews is significantly different (more predictable, less diverse / contentful) when compared to movie and restaurant reviews. We then turn to a final set of classification models that leverage transfer learning from the language models to see how predictive the texts can really be with respect to the associated binary outcomes. Finally, we report the results of estimating language models for the 4 auxiliary datasets introduced in the previous section.

The main conclusion of this extensive series of models is that the IMR corpus is an outlier, and it would be easy to make the IMR process fully automatic: it is pretty straightforward to train models that generate high-quality, realistic IMR reviews and generate binary decisions that are very reliably associated with these reviews. In contrast, movie and restaurant reviews produced by unpaid volunteers (as well as the 4 auxiliary datasets) exhibit more human-like depth, sophistication and attention to detail, so current NLP models do not perform as well on them.

3.1 Classification models

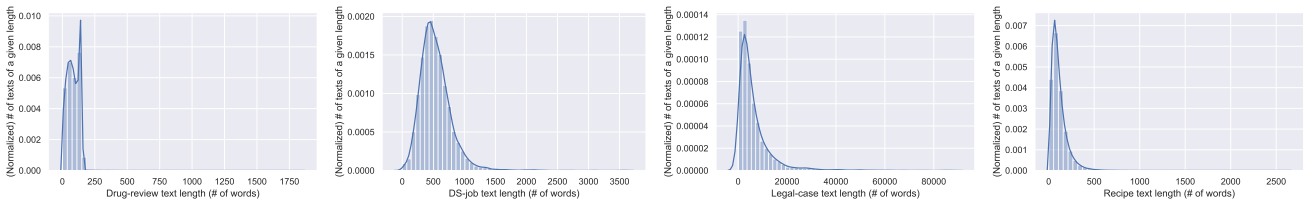
We regress outcomes (Upheld/Overturned for IMR or negative/positive sentiment for IMDB/Yelp) against the text of the corresponding findings / reviews. For the purposes of these basic classification models, as well as the topics models discussed in the following subsection, the texts were preprocessed as follows. First, we removed stop words; for the IMR dataset, we also removed the following high-frequency words: *patient*, *treatment*, *reviewer*, *request*, *medical* and *medically*, and for the IMDB dataset, we also removed the words *film* and *movie*. After part-of-speech tagging, we retained only nouns, adjectives, verbs and adverbs, since lexical meanings provide the most useful information for logistic (more generally, feed-forward) models and topic models. The resulting dictionary for the IMR dataset had 23,188 unique words. We ensured that the dictionaries for the IMDB and Yelp datasets were also between 23,000 and 24,000 words by eliminating infrequent words. Bounding the dictionaries for each dataset to a similar range helps mitigate dataset-specific modeling biases: having differently-sized vocabularies leads to differently-sized parameter spaces for the models.

We extracted features by converting each text into sparse bag-of-words vectors of dictionary length, which recorded how many times



(a) IMR (b) IMDB (c) Yelp

Figure 2: Histograms of text lengths (numbers of words per text) for the IMR, IMDB and Yelp corpora



(a) Drug Reviews (b) DS Jobs (c) Legal cases (d) Recipes

Figure 3: Histograms of text lengths (numbers of words per text) for the auxiliary datasets

each token occurred in the text. These feature representations were the input to all the classifier models we consider in this subsection. The multilayer perceptron model had a single hidden layer with 1,000 units and a ReLU non-linearity. The classification accuracies on the test data for all three datasets are provided in Table 3.

Table 3: Classification accuracy for basic models

	IMR	IMDB	Yelp
LOGISTIC REGRESSION	90.75%	86.30%	87.62%
MULTILAYER PERCEPTRON	90.94%	87.14%	88.92%

We see that the text of the findings / reviews is highly predictive of the associated binary outcomes, with the highest accuracy for the IMR dataset despite the fact that it contains half the observations of the other two data sets. We can therefore turn to a more in-depth analysis of the texts to understand what kind of textual justification is used to motivate the IMR binary decisions. To that end, we examine and compare the results of two unsupervised/self-supervised types of models: topic models and language models.

3.2 Topic models

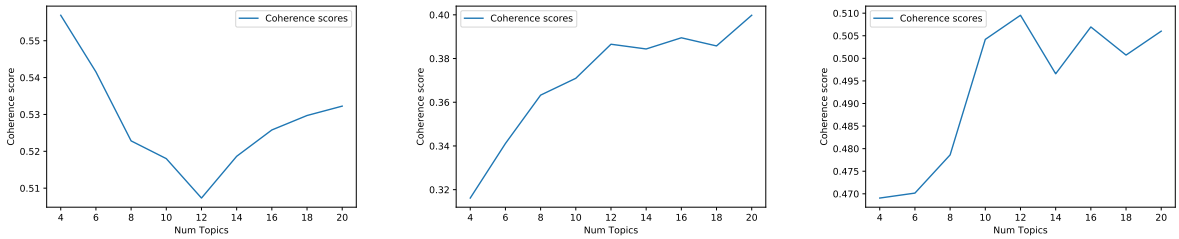
Topic modeling [17] is an unsupervised method that distills semantic properties of words and documents in a corpus in terms of probabilistic topics. The most widespread measure for topic model evaluation is the coherence score [14]. Typically, as we increase the number of topics from very few, say, 4 topics, to more of them, we see an increase in coherence score that tends to level out after a certain number of topics. When modeling the IMDB and Yelp datasets, we see exactly this behavior, as shown in Figure 4.

In contrast, the 4-topic model has the highest coherence score (0.56) for the IMR data set, also shown in Figure 4. Furthermore, as we add more topics, the coherence score drops. As the word clouds for the 4-topic model in Figure 5 show, these 4 topics mostly reflect the legalese associated with the IMR review procedure and very little, if anything, of the treatments and conditions that were the main point of the review. In contrast, the corresponding high-scoring topic models for the IMDB and Yelp datasets reflect actual features of movies, e.g., family-life movies, westerns, musicals etc., or breakfast/lunch places, restaurants, shops, bars, hotels etc.

Recall that IMRs are the legally-mandated last resort for patients seeking treatments (usually) ordered by their doctors, and which their health plan refuses to cover. The reviews are conducted exclusively based on documentation. Putting aside the fact that it is unclear how much effort is taken to ensure that the documentation is complete, especially for patients with extensive and complicated health records, we see that relatively little specific information about a patients' medical history, condition(s), or the recommended treatments are reflected in the text of these decisions. The text seems to consist largely of legalese about the IMR process, the health plan / providers, basic demographic information about the patient, and generalities about the medical service or therapy requested for the enrollee's condition.

3.3 Language models with transfer learning

Language models, specifically using neural networks, are usually recurrent-network or transformer based architectures designed to learn textual distributional patterns in an unsupervised or self-supervised manner. Recurrent-network models – on which we focus here – commonly use Long Short-Term Memory (LSTM) [7] "cells," which are able to learn long-term dependencies in sequences. Representing text as a sequence of words, language models build rich representations of the words, sentences, and their relations within a certain language. We estimate a language model for the IMR corpus using inductive sequential transfer learning, specifically ULMFiT [8]. Just as [8], we use the AWD-LSTM model [12], a vanilla



(a) IMR (b) IMDB (c) Yelp

Figure 4: Coherence scores for topic models (x-axis: number of topics; y-axis: coherence score)

Figure 5: Word clouds for the 4-topic IMR model

LSTM with 4 kinds of dropout regularization, embedding size of 400, 3 LSTM layers (1,150 units per layer), and a BPTT of size 70.

The AWD-LSTM model is pretrained on Wikitext-103 [13], consisting of 28, 595 preprocessed Wikipedia articles, with a total of 103 million words. This pretrained model is fairly simple (no attention, skip connections etc.), and the pretraining corpus is of modest size.

To obtain our final language models for the IMR, IMDB and Yelp corpora, we fine-tune the pretrained AWD-LSTM model using discriminative [18] and slanted triangular [8, 16] learning rates. We do the same kind of minimal text preprocessing as in [8].

The perplexity and categorical accuracy for the 3 language models are provided in Table 4. The perplexity for the IMR findings is much lower than for the IMDB / Yelp reviews, and the language model can correctly guess the next word more than half the time.

Table 4: Language-model perplexity and categ. accuracy

	IMR	IMDB	Yelp
PERPLEXITY	11.86	36.96	40.3
CATEGORICAL ACCURACY	53%	39%	29%

The IMR language model can generate high quality and largely coherent text, unlike the IMDB / Yelp models. Two samples of generated text are provided below (the 'seed' text is boldfaced).

- **The issue in this case is** whether the requested partial hospitalization program (PHP) services are medically necessary for treatment of the patient 's behavioral health condition . The American Psychiatric Association (APA) treatment guidelines for patients with eating disorders also consider PHP acute care to be the most appropriate setting for treatment , and suggest that patients should be treated in the least restrictive setting which is likely to be safe and effective . The PHP was initially recommended for patients who were based on their own medical needs , but who were
- **The patient was admitted** to a skilled nursing facility (SNF) on 12 / 10 / 04 . The submitted documentation states the patient was discharged from the hospital on 12 / 22 / 04 . The following day the patient 's vital signs were stable . The patient had been ambulating to the community with assistance with transfers , but has not had any recent medical or rehabilitation therapy . The patient had no new medical problems and was discharged in stable condition . The patient has requested reimbursement for the inpatient acute rehabilitation services provided

We see that the IMR language model is highly performant, despite the simple model architecture we used, the modest size of the pretraining corpus, and the small size of the IMR corpus. The quality of the generated text is also very high, particularly given all these limitations.

3.4 Classification with transfer learning

We further fine-tune the language models discussed in the previous subsection to train classifiers for the three datasets. Following [4, 8], we gradually unfreeze the classifier models to avoid catastrophic forgetting.

The results of evaluating the classifiers on the withheld test sets are provided in Table 5. Despite the fact that the IMR dataset contains half of the classification observations of the other two datasets, we obtain the highest level of accuracy when predicting binary Upheld/Overturned decisions based on the text of the IMR findings.

Table 5: Accuracy for transfer-learning classifiers

	IMR	IMDB	Yelp
CLASSIFICATION ACCURACY	97.12%	94.18%	96.16%

Table 6: Comparison of language models across all datasets. Best performing metrics are boldfaced.

Dataset	Perplexity	Categorical Accuracy
IMR reviews	**11.86**	**0.53**
Legal cases	18.17	0.43
DS Jobs	22.14	0.41
Drug reviews	25.06	0.36
Recipes	29.56	0.39
IMDB	36.96	0.39
Yelp	40.3	0.29

3.5 Models for auxiliary corpora

We also estimated topic and language models for the 4 auxiliary corpora (drug reviews, DS jobs, legal cases and cooking recipes). The associations between coherence scores and number of topics for these 4 corpora was similar to the ones plotted in Figure 4 above for the IMDB and Yelp corpora. For all 4 auxiliary corpora, the best topic models had at least 14 topics, often more, with coherence scores above 0.5. The quality of the topics was also high, with intuitively coherent and contentful topics (just like IMDB / Yelp).

The perplexity and accuracy of the ULMFiT language models on previously-withheld test data are provided in Table 6, which contains the results for all the 7 datasets under consideration in this paper. We see that the predictability of the IMR corpus, as reflected in its perplexity and categorical accuracy scores, is still clearly higher than the 4 auxiliary corpora. The perplexity of the legal-case corpus (18.17) is somewhat close to the IMR perplexity (11.86), but we should remember that the legal-case corpus is about 5 times larger than the IMR corpus. Furthermore, the legal-case categorical accuracy of 43% is still substantially lower than the IMR accuracy of 53%. Notably, even the recipe corpus, which is about 20 times larger than the IMR corpus ($\approx$ 117.5 vs. $\approx$ 5.5 million words) does not have test-set scores similar to the IMR scores.

The results for these 4 auxiliary corpora indicate that the IMR corpus is an outlier, with very highly templatic and generic texts.

4 DISCUSSION

The models discussed in the previous section show that language-model learning is significantly easier for IMRs compared to the other 6 corpora. As can be seen in Table 6, perplexity in the language model for IMR reviews is clearly lower than even legal cases, for which we expect highly templatic language and high similarity between texts. This pattern can be clearly observed in Figure 6, with the IMR corpus clearly at the very end of the high-to-low predictability spectrum.

One would not expect such highly predictable texts in an ideal scenario, where each medical review is thorough, and each decision is accompanied by strong medical reasoning relying on the specifics of the case at hand, and based on an objective physician's, or team of physicians', opinion as to what is in the patient's best interest. Arguably, these medically complex cases are as diverse as Hollywood blockbusters or fashionable restaurants – the patients themselves certainly experience them as unique and meaningful –, and their reviews should be similarly diverse, or at most as templatic as a job posting or a cooking recipe. We wouldn't expect

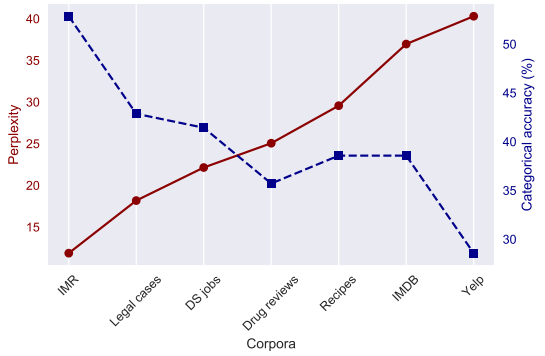


Figure 6: Comparison of language-model perplexity and categorical accuracy across all the datasets.

these medical reviews to be so much more predictable and generic than less socially consequential reviews of movies and restaurants.

What are the ethical and potentially legal consequences of these findings? First, while state legislators assume we have strong health-insurance related consumer protections in place, an image DMHC goes to great lengths to promote, we find the reviews to be upholding insurance plan denials at rates that exceed what one might expect, given that the treatments in question are frequently being ordered by a treating physician, and that the IMR process is the last stage in a bureaucratically laborious (hence high-attrition) process of appealing health-plan denials.

Second, given that the IMR process creates an implied relation of care between the reviewers hired by MAXIMUS and the patient – since reviewers are, after all, being entrusted with the best interests of the patient without regard to cost –, one can hardly say that they are fulfilling their obligations as doctors to their patient with such seemingly rote, perfunctory reviews.

Third, if IMR processes were designed to make sure that (*i*) treatment decisions are being made by doctors, not by profit-driven businesses, and (*ii*) insurance companies cannot welch on their responsibilities to plan members, one must wonder whether prescribing physicians are wrong more than half the time. Do American doctors really order so many erroneous, medically unnecessary treatments and medications? If so, how is it possible that they are so committed and confident in them that they are willing to escalate the appeal process all the way to the state-managed IMR stage? Or is it that IMRs often serve as a final rubber stamp for health-insurance plan denials, failing their stated mission of protecting a vulnerable population?

We end this discussion section by briefly reflecting on the way we used ML/NLP methods for social good problems in this paper. Overwhelmingly, the social-good applications of these methods and models seem to be predictive in nature: their goal is to improve the outcomes of a decision-making process, and the improvement is evaluated according to various performance-related metrics. An important class of metrics that are currently being developed have to do with ethical, or 'safe,' uses of ML/AI models.

In contrast, our use of ML models in this paper was analytical, with the goal of extracting insights from large datasets that enable us to empirically evaluate how well an established decision-making

process with high social impact functions. Data analysis of this kind, more akin to hypothesis testing than to predictive modeling, is in fact one of the original uses of statistical models / methods.

Unfortunately, using ML models in this way does not straightforwardly lead to plots showing how ML models obviously improve metrics like the efficiency or cost of a process. We think, however, that there are as many socially beneficial opportunities for this kind of data-analysis use of ML modeling as there are for its predictive uses. The main difference between them seems to be that the data-analysis uses do not lead to more-or-less immediately measurable products. Instead, they are meant to become part of a larger argument and evaluation of a socially and politically relevant issue, e.g., the ethical status of current health-insurance related practices and consumer protections discussed here. What counts as 'success' when ML models are deployed in this way is less immediate, but could provide at least as much social good in the long run.

5 CONCLUSION AND FUTURE WORK

We examined a database of 26,361 IMRs handled by the California DMHC through a private contractor. IMR processes are meant to provide protection for patients whose doctors prescribe treatments that are denied by their health insurance.

We found that, in a majority of cases, IMRs uphold the health insurance denial, despite DMHC's claim to the contrary. In addition, we analyzed the text of the reviews and compared them with a sample of 50,000 Yelp reviews and the IMDB movie review corpus. Despite the fact that these corpora are basically twice as large, we can construct a very good language model for the IMR corpus, as measured by the quality of text generation, as well as its low perplexity and high categorical accuracy on unseen test data. These results indicate that movie and restaurant reviews exhibit a much larger variety, more contentful discussion, and greater attention to detail compared to IMR reviews, which seem highly templatic and perfunctory in comparison. We see similar trends in topic models and classification models predicting binary IMR outcomes and binarized sentiment for Yelp and IMDB reviews.

These results were further confirmed by topic and language models for four other specialized-register corpora (drug reviews, data science job postings, legal-case reports and cooking recipes).

We are in the process of extending our datasets with (*i*) workers' comp cases from California and (*ii*) private insurance cases from other states. This will enable us to investigate if the reviews for workers' comp cases are substantially different from the DMHC IMR data (the percentage of upheld decisions is much higher for workers' comp: $\approx 90\%$), as well as if the reviews vary substantially across states.

Another direction for future work is to follow up on our preliminary qualitative research with a survey of patients that have experienced the IMR process to see if these patients agree with the DMHC-promoted message that the IMR process provides strong consumer protection against unjustified health-plan denials. This could also enable us to verify if the medical documentation collected during the IMR process is complete and actually taken into account when the decision is made.

The ultimate upshot of this project would be a list of recommendations for the improvement of the IMR process, including but not limited to (*i*) adding ways for patients to check that all the relevant documentation has been collected and will be reviewed, and (*ii*) identifying ways to hold the anonymous reviewers to higher standards of doctor-patient care.

ACKNOWLEDGMENTS

We are grateful to four KDD-KiML anonymous reviewers for their comments on an earlier version of this paper. We gratefully acknowledge the support of the NVIDIA Corporation with the donation of two Titan V GPUs used for this research, as well as the UCSC Office of Research and The Humanities Institute for a matching grant to purchase additional hardware. The usual disclaimers apply.

REFERENCES

[1] Leatrice Berman-Sandler. 2004. Independent Medical Review: Expanding Legal Remedies to Achieve Managed Care Accountability. *Annals Health Law* 13 (2004).
[2] Kenneth H. Chuang, Wade M. Aubry, and R. Adams Dudley. 2004. Independent Medical Review Of Health Plan Coverage Denials: Early Trends. *Health Affairs* 23, 6 (2004), 163–169. https://doi.org/10.1377/hlthaff.23.6.163
[3] Angus Deaton and Nancy Cartwright. 2018. Understanding and misunderstanding randomized controlled trials. *Social Science and Medicine* 210 (2018), 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005
[4] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 1615–1625. https://doi.org/10.18653/v1/D17-1169
[5] Filippo Galgani and Achim Hoffmann. 2011. LEXA: Towards Automatic Legal Citation Classification. In *AI 2010: Advances in Artificial Intelligence*, Jiuyong Li (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 445–454.
[6] Felix Gräundefineder, Surya Kallumadi, Hagen Malberg, and Sebastian Zaunseder. 2018. Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning *(DH '18)*. Association for Computing Machinery, New York, NY, USA, 121–125. https://doi.org/10.1145/3194658.3194677
[7] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (Nov. 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
[8] Jeremy Howard and Sebastian Ruder. 2018. Fine-tuned Language Models for Text Classification. *CoRR* abs/1801.06146 (2018). arXiv:1801.06146 http://arxiv.org/abs/1801.06146
[9] Shanshan Lu. 2018. Data Scientist Job Market in the U.S. https://www.kaggle.com/sl6149/data-scientist-job-market-in-the-us More info available here: https://github.com/Silvialss/projects/tree/master/IndeedWebScraping.
[10] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis *(HLT '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 142–150.
[11] Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. 2019. Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images. *IEEE Trans. Pattern Anal. Mach. Intell.* (2019).
[12] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2017. Regularizing and Optimizing LSTM Language Models. *CoRR* abs/1708.02182 (2017).
[13] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer Sentinel Mixture Models. *CoRR* abs/1609.07843 (2017).
[14] Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the Space of Topic Coherence Measures *(WSDM '15)*. ACM, New York, NY, USA, 399–408. https://doi.org/10.1145/2684822.2685324
[15] Shirley Eiko Sanematsu. 2001. Taking a broader view of treatment disputes beyond managed care: Are recent legislative efforts the cure? *UCLA Law Review* 48 (2001).
[16] Leslie N. Smith. 2017. Cyclical learning rates for training neural networks. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*. IEEE. 464–472.
[17] Mark Steyvers and Tom Griffiths. 2007. *Probabilistic Topic Models*. Lawrence Erlbaum Associates.
[18] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks?. In *Advances in Neural Information Processing Systems*. 3320–3328.
[19] Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. *CoRR* abs/1509.01626 (2015). arXiv:1509.01626 http://arxiv.org/abs/1509.01626