Textual Evidence for the Perfunctoriness of Independent Medical Reviews

Adrian Brasoveanu[†], Megan Moodie[†], Rakshit Agrawal[‡]
[†]UC Santa Cruz, [‡]Camio Inc.

KDD KiML Workshop · August 2020

Origin of IMRs

Independent Medical Review (IMR) processes

- IMRs are meant to provide protection for patients whose doctors prescribe treatments that are denied by their health insurance
  - private or workers' comp insurance
  - we focus exclusively on privately insured patients here
- Laws requiring IMRs were established in California and other states in the late 1990s
  - patients and their doctors were concerned that health insurance plans deny coverage for medically necessary services to maximize profit
- IMRs are regularly used to settle disputes between patients and their health insurers over what is medically necessary or experimental/investigational care (Berman-Sandler, 2004)

Medical necessity and Utilization Review

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
  - course of treatment for a specific condition

Utilization Review (UR)

- UR is the oversight mechanism through which private insurers control costs
- services rendered by a health care provider are reviewed to determine whether the services are medically necessary
- Services that are not deemed medically necessary or fall outside the contractual terms of the insurance plan are not covered

Experimental or investigational care

Experimental or investigational procedures / treatments

- the health plan (but not the patient's doctor) considers them non-routine medical care, or takes them to be scientifically unproven
- typical argument: evidence for treatment effectiveness fails the prevailing standard of scientific evidence, usually randomized control trials (RCTs)
- experimental/investigational treatments that get denied include promising treatments that have not been fully tested in clinical RCTs because
  - they are new
  - they treat rare conditions, so RCT costs (if RCTs are feasible at all) won't be recovered

Randomized Control Trials (RCTs)

- expensive & time-consuming
- run by pharmaceutical companies only if the treatment is ultimately estimated to be profitable

> *[RCTs] 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.[. . .]*
> *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.*
> *(Deaton and Cartwright, 2018)*

Inflexibly applying the RCT 'gold standard' (e.g., to conditions that are less prevalent): ignoring a doctor's recommendation in a seemingly well-reasoned & scientific way.

IMRs in the medical review timeline

IMRs are the third and final stage in the medical review process:

- after in-person and possibly repeated examination of the patient, the doctor recommends a treatment, 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
- this triggers a second stage of reviews by the health-insurance provider, during which
  - a patient can supply additional information
  - a doctor may engage in "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
  - per California law, an IMR grievance form (and some additional information) is included with the denial letter

The patient–reviewer relationship

- IMRs must be initiated by patient & submitted to the California Department of Managed Health Care (DMHC; 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
- in theory, the IMR creates a new relationship of care between the patient and the reviewing physician(s) hired by a private contractor on behalf of DMHC
  - 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

Road map

Main argument and predictions

- IMRs are the final stage in a long bureaucratic process, so specifics of patient history and recommended treatments are likely crucial

Therefore, we expect:

- the text of the IMRs, which justifies the final determination, to be highly individualized
  - justification for the final decision should involve the particulars of the treatment and the patient's medical history and conditions
- a reasoned, thoughtful IMR to not be highly generic and templatic
  - 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
- high inter-IMR similarity and 'templaticity' only if they are reduced to a more or less automatic application of some prespecified set of rules (rubber-stamping)

Main results and their limits

- 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, we might have been able to provide much stronger evidence that:
  - IMRs should have a significantly higher percentage of overturns
  - the IMR process should be improved in various ways, e.g.,
    - patients should be able to check that all the relevant documentation has been collected and will be reviewed
    - 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 not likely to become available in the (near) future

Main results and their limits (ctd.)

Main dataset: the corpus of IMR decisions made available by the California DMHC site as of June 2019 (a total of 26,631 cases spanning the years 2001-2019)

A qualitative inspection 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
  - very little consideration of the prevalence of the underlying condition(s)
  - no 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)

Main results and their limits (ctd.)

Our main goal here:

- investigate to what extent NLP/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

Specifically:

- We analyze the text of the IMR reviews and compare them with a sample of 50,000 Yelp reviews (Zhang et al., 2015) and the corpus of 50,000 IMDB movie reviews (Maas et al., 2011)
- As the size of data has significant consequences for language-model training (and other types of NLP/ML models), we expect models trained on the Yelp and IMDB corpora to outperform models trained on the IMR corpus
  - the IMDB corpus is twice as large as the IMR corpus
  - the Yelp samples contain almost twice as many reviews

Main results and their limits (ctd.)

- However, we are able to construct a very good language model for the IMR corpus using inductive sequential transfer learning, specifically ULMFiT (Howard and Ruder, 2018)
- 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 (Yelp) and 37 (IMDB)
  - categorical accuracy: 0.29 (Yelp) and 0.39 (IMDB)
- We see similar trends in topic models (Steyvers and Griffiths, 2007), and also classification models predicting binary IMR outcomes and binarized sentiment for Yelp and IMDB reviews
- Movie and restaurant reviews exhibit a much larger variety, more contentful discussion, and greater attention to detail compared to IMR reviews

Main results and their limits (ctd.)

- To mitigate potentially significant register differences between IMRs and movie or restaurant reviews, we examine four additional corpora:
  - drug reviews (Gräundefineder et al., 2018)
  - data science job postings (Lu, 2018)
  - legal case summaries (Galgani and Hoffmann, 2011)
  - cooking recipes (Marin et al., 2019)
- These specialized-register corpora are potentially more similar to IMRs than IMDB or Yelp:
  - more likely to be highly similar
  - include boilerplate text
  - 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

Four kinds of models

- All datasets split into training (80%), validation (10%) and test (10%) sets
- Test sets were only used for the final model evaluation
- We discuss four kinds of models:
  1. classification models
  2. topic models
  3. language models with transfer learning
  4. classification models with transfer learning

Classification models

- We regress outcomes (Upheld/Overturned for IMR or negative/positive sentiment for IMDB/Yelp) against the text of the findings / reviews
- We extract features by converting each text into sparse bag-of-words vectors of dictionary length
- Multilayer perceptron: one hidden layer with 1,000 units & ReLU non-linearity

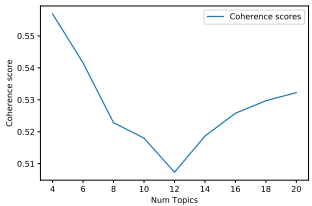
Table: Classification accuracy for basic models

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

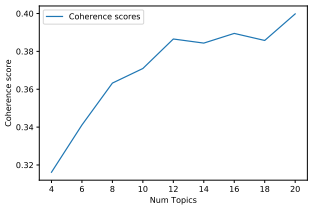
- Text of findings/reviews is highly predictive of the associated binary outcomes
- Highest accuracy for the IMR dataset despite the fact that it contains half the observations of the other two data sets

Topic models

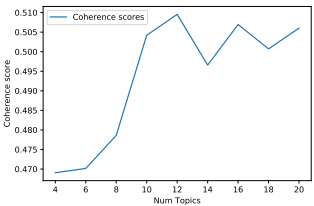
- Topic modeling (Steyvers and Griffiths, 2007) distills semantic properties of words and documents in a corpus in terms of probabilistic topics
- Topic model evaluation: coherence score (Röder et al., 2015)
- As we increase the number of topics from very few (e.g., 4) topics to more of them: increase in coherence score that levels out (IMDB and Yelp)
- The 4-topic model has the highest coherence score (0.56) for the IMR data set; as we add more topics, the coherence score drops



IMR



IMDB



Yelp

Topic models (ctd.)

- Word clouds for the 4-topic IMR model mostly reflect the legalese associated with the IMR review procedure, and very little of the treatments and conditions under review
- High-scoring IMDB and Yelp models reflect contentful features: family-life movies, westerns, musicals, or breakfast/lunch places, restaurants, shops, bars, hotels etc.

Language models with transfer learning

Language models, specifically using neural networks:

- usually recurrent-network (our focus here) or transformer based architectures
- designed to learn textual distributional patterns in a self-supervised manner
- recurrent-network models commonly use Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) “cells”
  - able to learn long-term dependencies in sequences

Language models with transfer learning

We estimate a language model for the IMR corpus using inductive sequential transfer learning (ULMFiT, Howard and Ruder 2018):

- Just as Howard and Ruder (2018), we use the AWD-LSTM model (Merity et al., 2017b)
  - a vanilla 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 (Merity et al., 2017a)
  - $28,595$ 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

Lang. models with transfer learning (ctd.)

To obtain our final language models for the IMR, IMDB and Yelp corpora:

- we fine-tune the pretrained AWD-LSTM model, using discriminative (Yosinski et al., 2014) and slanted triangular (Howard and Ruder, 2018; Smith, 2017) learning rates

Table: Language-model perplexity and categ. accuracy

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

- 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
- the quality of the generated text is also very high (unlike for IMDB / Yelp)

Lang. models with transfer learning (ctd.)

IMR language model generates high quality coherent text ('seed' text 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
- **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

Classification models with transfer learning

We further fine-tune the language models to train classifiers for the three datasets:

- we gradually unfreeze the classifier models to avoid catastrophic forgetting (Felbo et al., 2017; Howard and Ruder, 2018)

Table: Accuracy for transfer-learning classifiers

	IMR	**IMDB**	**Yelp**
CLASSIFICATION ACCURACY	97.12%	94.18%	96.16%

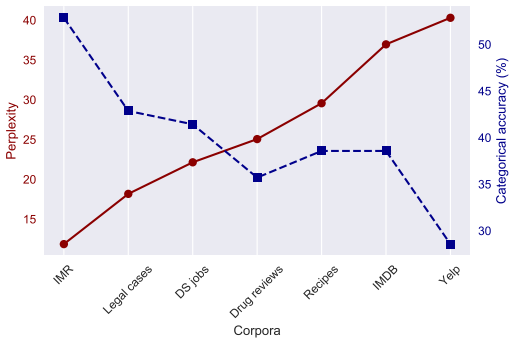
We obtain the highest level of accuracy when predicting binary Upheld/Overturned decisions based on the text of the IMR findings

- despite the fact that the IMR dataset contains half of the classification observations of the other two datasets

Language models for auxiliary datasets

The predictability of the IMR corpus – as reflected in its perplexity and categorical accuracy scores – is also clearly higher than the 4 auxiliary corpora.

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



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

Lang. models for auxiliary datasets (ctd.)

- Perplexity of the legal-case corpus (18.17) is somewhat close to the IMR perplexity (11.86)
- . . . but the legal-case corpus is about 5 times larger than the IMR corpus
- Legal-case categorical accuracy of 43% is still substantially lower than the IMR accuracy of 53%
- 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 IMR

Conclusion

- Language-model learning is significantly easier for IMRs compared to the other 6 corpora
  - the IMR corpus is at the very end of the high-to-low predictability spectrum
- We would not expect such highly predictable texts if each decision was accompanied by thorough reasoning relying on the specifics of the case
  - these medically complex cases are arguably as diverse as Hollywood blockbusters or fashionable restaurants
  - the patients themselves certainly experience them as unique and meaningful
  - their reviews should be similarly diverse, or at most as templatic as a job posting or a cooking recipe

Conclusion (ctd.)

Ultimate upshot of this work: a list of recommendations for the improvement of the IMR process, including

1. adding ways for patients to check that all the relevant documentation has been collected and will be reviewed
2. identifying ways to hold the anonymous reviewers to higher standards of doctor-patient care

Conclusion (ctd.)

Final note about the way we used NLP/ML methods for social good problems:

- 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
  - the improvement is evaluated according to various performance-related metrics
  - important class of metrics currently being developed: ethical, or 'safe,' uses of ML/AI models
- In contrast, our use of ML models in this paper was analytical:
  - goal: extracting insights from large datasets to empirically evaluate an established decision-making process with high social impact
- Our use is more akin to data summarization and hypothesis testing (the other two uses of statistical methods & models) than predictive modeling

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.

References I

Berman-Sandler, Leatrice (2004). “Independent Medical Review: Expanding Legal Remedies to Achieve Managed Care Accountability”. In: *Annals Health Law* 13.

Deaton, Angus and Nancy Cartwright (2018). “Understanding and misunderstanding randomized controlled trials”. In: *Social Science and Medicine* 210, pp. 2–21. DOI: 10.1016/j.socscimed.2017.12.005.

Felbo, Bjarke et al. (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*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 1615–1625. DOI: 10.18653/v1/D17-1169.

Galgani, Filippo and Achim Hoffmann (2011). “LEXA: Towards Automatic Legal Citation Classification”. In: *AI 2010: Advances in Artificial Intelligence*. Ed. by Jiuyong Li. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 445–454.

Gräundefineder, Felix et al. (2018). “Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning”. In: *Proceedings of the 2018 International Conference on Digital Health*. DH ’18. Lyon, France: Association for Computing Machinery, pp. 121–125. ISBN: 9781450364935. DOI: 10.1145/3194658.3194677. URL: https://doi-org.stanford.idm.oclc.org/10.1145/3194658.3194677.

References II

- Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1997). “Long Short-Term Memory”. In: *Neural Comput.* 9.8, pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. URL: http://dx.doi.org/10.1162/neco.1997.9.8.1735.
- Howard, Jeremy and Sebastian Ruder (2018). “Fine-tuned Language Models for Text Classification”. In: *CoRR* abs/1801.06146. arXiv: 1801.06146. URL: http://arxiv.org/abs/1801.06146.
- Lu, Shanshan (2018). *Data Scientist Job Market in the U.S.* More info available here: https://github.com/Silvialss/projects/tree/master/IndeedWebScraping. URL: https://www.kaggle.com/sl6149/data-scientist-job-market-in-the-us.
- Maas, Andrew L. et al. (2011). “Learning Word Vectors for Sentiment Analysis”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. HLT ’11. Portland, Oregon: Association for Computational Linguistics, pp. 142–150. ISBN: 978-1-932432-87-9.
- Marin, Javier et al. (2019). “Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images”. In: *IEEE Trans. Pattern Anal. Mach. Intell.*
- Merity, Stephen et al. (2017a). “Pointer Sentinel Mixture Models”. In: *CoRR* abs/1609.07843.
- Merity, Stephen et al. (2017b). “Regularizing and Optimizing LSTM Language Models”. In: *CoRR* abs/1708.02182.

References III

- Röder, Michael et al. (2015). “Exploring the Space of Topic Coherence Measures”. In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. WSDM ’15. Shanghai, China: ACM, pp. 399–408. ISBN: 978-1-4503-3317-7. DOI: 10.1145/2684822.2685324.
- Smith, Leslie N. (2017). “Cyclical learning rates for training neural networks”. In: *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on. IEEE*, pp. 464–472.
- Steyvers, Mark and Tom Griffiths (2007). “Probabilistic Topic Models”. In: *Handbook of Latent Semantic Analysis*. Ed. by T. Landauer et al. Lawrence Erlbaum Associates. ISBN: 1410615340.
- Yosinski, Jason et al. (2014). “How transferable are features in deep neural networks?” In: *Advances in Neural Information Processing Systems*, pp. 3320–3328.
- Zhang, Xiang et al. (2015). “Character-level Convolutional Networks for Text Classification”. In: *CoRR* abs/1509.01626. arXiv: 1509.01626. URL: http://arxiv.org/abs/1509.01626.