An investigation into the use of artificial intelligence in regulatory decision making about complaints about nurses in the US, UK and Australia

Robert Jago, Anna van der Gaag, Kostas Stathis,
Ivan Petej, Piyawat Lertvittayakumjorn, Yamuna Krishnamurthy, Yang Gao,
Juan Caceres Silva, Michelle Webster,
Royal Holloway University of London UK
Ann Gallagher, University of Exeter UK
Zubin Austin, University of Toronto, Canada

Correspondence concerning this article should be addressed to Robert Jago, Royal Holloway University of London, UK. Email: Robert.jago@rhul.ac.uk

Word count: 3,475

Funding
Funding was provided by the Centre for Regulatory Excellence National Council of State Boards of Nursing (NCSBN).

Conflicts of Interest:
None

This manuscript is not under review by another publication and will not be sent to any other publishers whilst under review by JNR, and has not previously been published elsewhere, even in another language.

Keywords: Complaint resolution, artificial intelligence, investigation, nursing discipline, nursing regulation

Abstract

This project aimed to develop an Artificial Intelligence (AI) based tool for improving the consistency and efficiency of decision-making in the nursing complaints process in three jurisdictions. Its primary focus was on improving the processing of complaints at the screening stage of investigation. It was not designed to replace human judgement but to provide staff with three data-driven decision support tasks: an independent risk classification of the case, a comparison with previous similar cases and a cross reference to relevant parts of the regulatory standards or rules in each jurisdiction. Three nurse regulators in the United States, the United Kingdom and Australia provided anonymized data from 5,700 cases for tool design and testing. Regulatory staff were involved in each stage of development. The consensus across the regulators was that this tool has significant potential to improve the efficiency of decision-making in disciplinary processes in nurse regulation nationally and internationally.

Introduction

Artificial Intelligence (AI) tools are increasingly being utilized to improve the quality and speed of processing large scale data sets in commercial and public sector organisations worldwide (Cam et. al., 2019; Susskind, 2020). The Covid-19 pandemic has accelerated this process (WEC, 2020), leading to increased levels of automation and the demand for new technical skills in the workforce. It has been suggested that AI has the potential to make humans more productive across many sectors, if society takes a human centric approach to technological advances (Acemoglu and Restrepo, 2020).

There is increasing evidence that AI models achieve human-like performance in a variety of settings. For example, cancer diagnosis using AI has reached high levels of agreement with human specialists (McKinney et. al., 2020). Emerging results suggest that combinations of human judgement and machine learning platforms may increase validity and fairness when compared with human judgement alone. In the legal arena, AI tools are being used to create summaries from case documents (Waltl et. al., 2017), evaluate the impact of a ruling on future rulings and classify court cases using processes that model human searches (Leibon et. al., 2016). In human resource management, AI tools have been developed to help detect evidence of harassment in emails (Sulea et. al., 2017; Woodford, 2020). These types of disruptive innovations have demonstrated positive impacts in a wide variety of sectors but to date they have

rarely been tested in a health regulatory environment. In Australia, (Spittal et. al., 2019) have developed algorithms to assess risk factors in health professionals, such as medical specialty and occurrence of previous complaints. However, these tools have not, to our knowledge, been used as decision support tools in health disciplinary functions.

Disciplinary decision making is a complex process reliant upon multiple sources of evidence and in-depth understanding of rules. (Cronquist, 2013) explored nurse complaint and disciplinary process within the United States (US) from the screening of complaints through investigation, prosecution, Board of Nursing (BON) actions, and compliance monitoring. Whilst the vast majority of nurses practice in a safe and competent manner (Cronquist, 2013), any departure from such competencies may result in a report or complaint being filed against them. Within the US, such complaints may be submitted by patients, family members, employers, co-workers, nurses and other professionals. The initial step of the allegation review is primarily a manual process which requires significant human, financial and technological resources.

Recent years have seen a rapid evolution in healthcare design and delivery with expanded scope and complexity of nursing practice. As the scope of nursing practice evolves, the workload for regulatory staff, responsible for reviewing the incoming complaints against nurses, is also likely to increase (Sanson, 2017). A report by the United Kingdom's (UK) medical regulator on the activities of nine professional health regulators in the UK identified a 32% increase in complaints against health practitioners over the preceding six years (GMC, 2017). Analysis has shown that a large proportion of these complaints could be described as 'low' risk complaints, because they are not upheld and there is no evidence of harm to patients or their families (NMC, 2020).

Literature review

AI is often understood as the scientific and engineering effort to make machines intelligent, by building them with capabilities traditionally reserved for humans, such as using language, forming abstractions, solving problems and learning from experience. In this context, machine learning usually refers to a set of trained models working in tandem to process observational data and produce outputs of value. These models are typically mathematical and unveil regularities from data (Bishop, 2006). Well known applications involve data classification, data summarization, estimation of relationships

between variables and generation of models that fit observed data (Shalev-Shwartz et. al., 2014). A broader family of machine learning methods based on artificial neural networks, known as deep learning, have become increasingly popular recently in recognition tasks, such as Natural Language Processing (NLP) (Devlin et. al., 2019), which is key to this work.

AI in Healthcare. Recent technological advances and the abundance of new data have contributed to a rapid increase in the development of machine learning applications within clinical decision-support systems. These systems were designed to assist and improve the workloads of healthcare practitioners, and they have been applied to tasks such as clinical diagnostics and selection of patients for clinical trials (Assale et. al, 2019; Davenport and Kalakota, 2019; Brooks, 2019). The NCSBN's Environmental Scan Report in 2020 suggested uses of AI in areas such as health screening and diagnostics, AI enabled automated processes and AI assisted patient engagement are growing rapidly, and will have legal and ethical implications for regulators (NCSBN, 2020) While the use of AI technologies within healthcare to date shows great promise, thus far it has not been tested in a nurse regulatory environment.

AI and Ethics. The advent of AI has brought with it a wide variety of responses from policy makers and practitioners. Many of the strongest objections to the development of these tools stem from concerns about privacy, fairness and transparency, and human rights (Benton et. al., 2020). For example, (Gianfrancesco, 2018) showed that AI systems applied within clinical decision support can potentially exhibit important societal biases, and if used incorrectly, can amplify healthcare disparities. (Obermeier et. al., 2019) reported that a machine learning algorithm used by many US healthcare insurers incorporated a faulty metric to determine which patients were high-risk and qualified for additional care management. AI algorithms used in other fields, such as law enforcement, academic settings and marketing, have also been found to exhibit some degree of implicit bias (Cossins, 2018; Levin, 2019). In relation to transparency, healthcare regulators are also often challenged by the perceived obscurity of the AI decision-making process (Davenport and Kalakota, 2019). Regarding accountability, (Kent 2019) argued that, as future AI applications will inevitably make errors, there is a strong need for discipline or systems-based responses to be in place when errors occur, to ensure patient safety.

Governing AI. In response to these concerns over the ethics of AI systems, researchers have called for system-wide guidance and codes of practice to ensure that AI development complies with ethical principles (Babuta, et. al., 2018; Sharkey, 2019). Few would argue against the need for rigorous governance arrangements and compliance with the highest standards of data protection in the development of AI tools. There have been examples of legal and governance failures that have created distrust in AI across the world (Forbes, 2020). In the US, legislators have proposed the Algorithmic Accountability Act (2019), which requires companies using high risk automated decision support systems to conduct algorithmic impact assessments. In Europe, GDPR legislation has been welcomed as a means of enforcing principles of fairness and transparency in relation to data storage and use. For AI, this includes identifying biases as part of product design, and distinguishing between non-interpretable ('black box") algorithms that should be subject to particularly high levels of testing and ongoing scrutiny (Vayena et. al, 2018) and interpretable algorithms, where the models provide insight about the inferences made about the data (Murdoch et. al., 2019). (Babuta et. al, 2018) observe that black box algorithms digest large data sets without being able to demonstrate their workings, and those from whom the data is derived have no knowledge of the decisions that have been made about them with the help of an algorithm. They call for system-wide guidance and codes of practice to ensure that AI development and deployment complies with ethical principles (Babuta et. al, 2018), including technical transparency and specifications about the availability of source code.

In designing our system, we were aware that the human consequences of a complaint can be far reaching for the individual, and their family and wider community. Our goal was therefore to design a system that was as accurate, transparent, unbiased and accountable as existing processes with the potential to improve these processes in terms of time, efficiency and confidence. The following section will describe the methods we used to achieve this goal and which we hope will be helpful to researchers designing similar systems in the future.

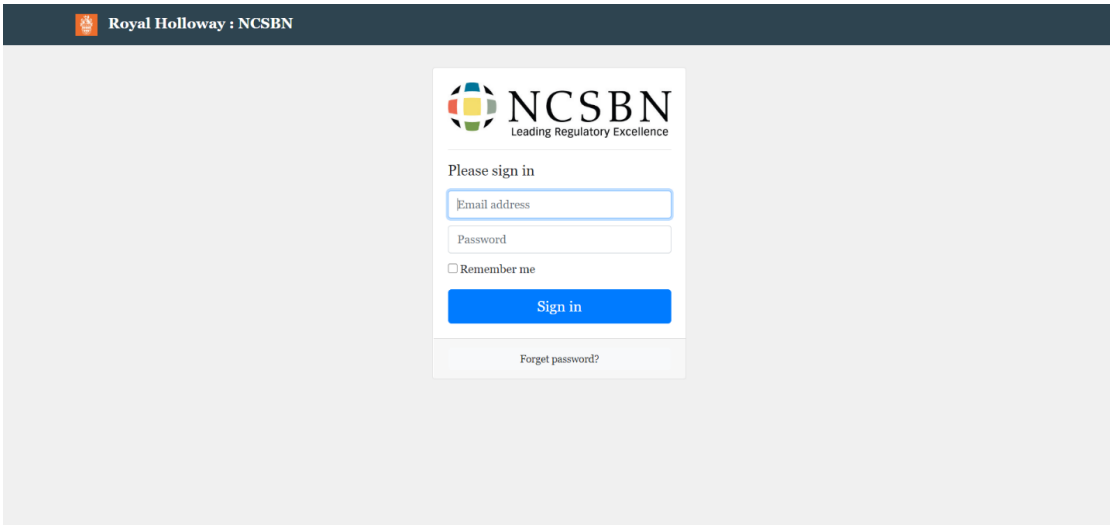
Methods

Nurse regulators in three jurisdictions agreed to participate in the research. All three shared the same aspiration to explore data driven solutions to the challenges of processing high volumes of complaints. Initially, we used open access complaints data from the financial sector as part of our preliminary modeling work. This allowed testing

of various combinations of approaches prior to using nurse complaints data to develop the prototype.

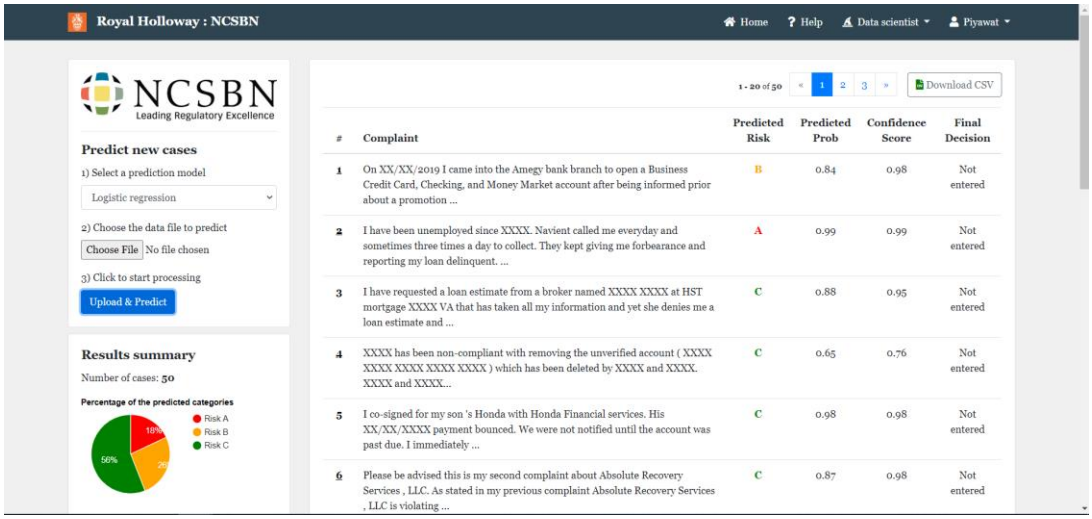
We developed a web-based application, which users in the three sites could access via a password protected portal (see Figure 1). This was the most appropriate design choice because the web application was easy to update and used a central high-performance server to process new complaints. This allowed case managers to upload their case files to a web server.

]

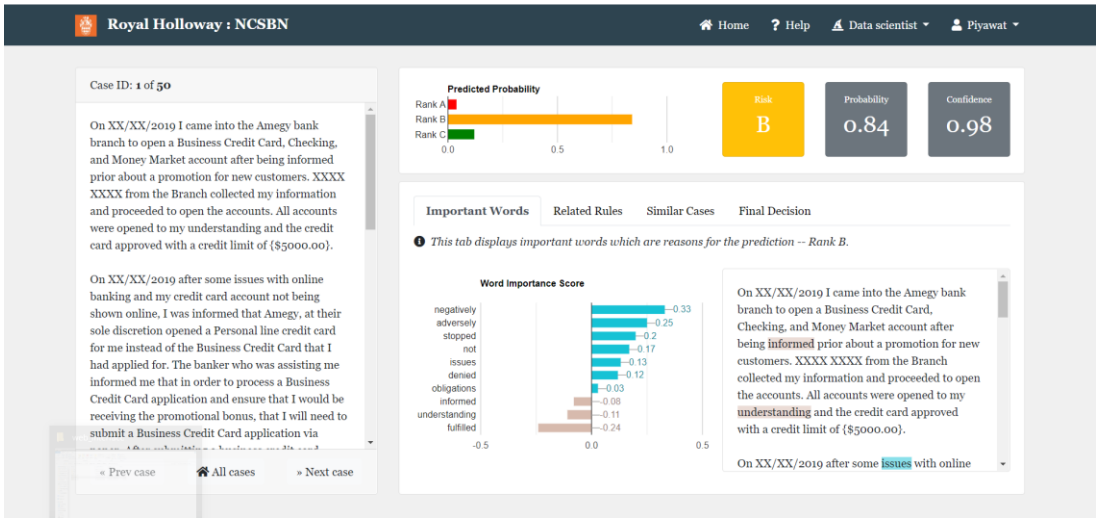


Insert Figure 1 here: The first page of the tool asking users to sign in

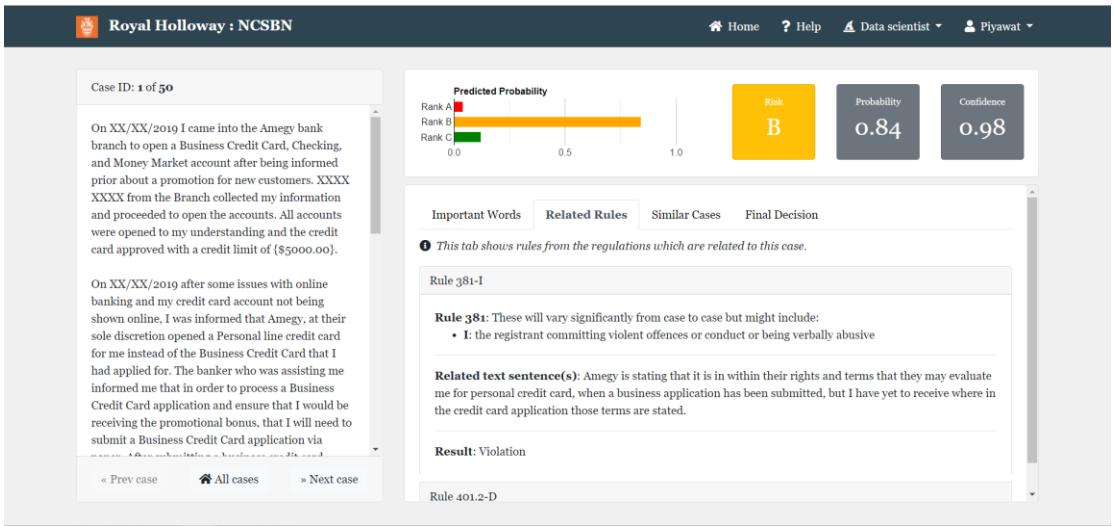
Due to privacy concerns with the data, Figures 2-4 give examples of outputs from the prototype using open access financial data. There are two main pages with which users interact after uploading a data file. Figure 2 depicts a table of all the uploaded cases and charts summarizing the statistics of the predictions. By clicking a row in the table, the users are redirected to a result page (Figure 3-4) for the specific case, showing outputs of the system including the predicted risk level, predicted probability, confidence scores, important words, similar cases, and relevant rules. In addition, users can provide feedback in response to the system outputs, recording their reasons for agreement or disagreement with the tool's risk assessment. The main reporting page provides a full summary of the results of all cases that have been uploaded.



Insert Figure 2 here : All uploaded cases, charts summaries and risk-level predictions.



Insert Figure 3 here: A specific case with the ratings of word importance scores



Insert Figure 4 here: A specific case with the relevant rules shown

The core features of the tool are described below.

Risk Prediction. The tool classifies the case as high risk or low risk and provides the probability of the risk prediction. This classification is achieved by using a *supervised machine learning* approach, where prominent characteristics of high-risk and low-risk cases are learnt from past cases that were processed manually and are referred to as *training data*. We used a technique called *ensemble learning* which created several machine learning models (i.e., base models) and combined their predictions to provide the final output (Wolpert, 1992). The advantage of this technique is that it combines complementary strengths of the base models to enhance the system accuracy. This also allows base models to learn from different parts of the data and then combine the results.

We also addressed gender bias in the training data by applying a technique called *gender swapping* (Zhao et. al., 2018). Previous research has shown that this technique is effective in mitigating gender biases in tasks such as abusive language detection (Sun et. al., 2019; Park et. al., 2018). In addition, the system highlighted words in the complaint which were considered important for predicting risk by the ensemble method in order to help case managers efficiently assess the case and verify or reject the prediction (see Figure 3).

Similar Cases Retrieval. We hypothesized that it would be helpful to link past complaints which were semantically similar to the new complaint so as to help case managers cross-check with previous judgements and improve consistency in decision making. First, we used a word-level similarity technique, called TF-IDF (Salton, 1988), to shortlist past complaints based on the overlapping of prominent words (Tata and Patel, 2007). Then, we fine-tuned the deep learning model BERT (Devlin et. al., 2019), and used it to return the top three cases to users (with similarity scores and the associated risk levels assessed by case managers in the past).

Relevant Standards Matching. We also aimed to link standards or rules from regulatory codes that were relevant to the new complaint to provide more information to the case managers. However, as the codes used by each regulator were different in terms of number, structure, and the applicable nursing roles, we designed this feature specifically for each of the three jurisdictions. The approaches we used relied on semantic text similarity (Reimers and Gurevych, 2019) and textual inference (Williams et. al., 2018) relating parts of the complaint to rules. This process identified the three most relevant rules to the case under consideration (see Figure 4).

User Feedback. In order to improve the system over time, we collected feedback from the users throughout the development of the tool and used this to design and fine tune the three features described above. Case managers provided feedback on risk classification of the new case, similarity to other cases, and the relevance of the rules to the given case. This feedback was especially important for identifying features of similar cases and relevant rules since there was insufficient training data to build supervised machine learning models for them.

Results

Datasets and Risk Prediction Accuracy

Table 1 provides an overview of the data and risk prediction accuracy for each jurisdiction. The system predicts whether each case is of low or high risk. The accuracy refers to the percentage of correct risk predictions for the test cases, while the baseline is the accuracy when all cases are predicted with the same majority risk in the training data.

The UK nurse regulator, the Nursing and Midwifery Council (NMC) provided 1,250 cases. Each case contained the redacted text of the allegation and additional non-textual data (including binary and categorical features) that provided further context for the case. The NMC redacted text data by replacing person names, organization names, dates, etc. with the tokens [PERSON], [ORGANISATION], [DATE], etc., so the resulting text was, more or less, grammatically correct and understandable. This resulted in a prediction accuracy of 71% (9% above baseline). Bias testing on the NMC data showed that the model did not use gender in making its decisions, probably due to the redacted nature of the text. The gender swapping technique can further reduce the bias in the model while sacrificing the accuracy by less than 1%.

The Australian Health Practitioner Agency (AHPRA) provided 1,300 cases with only the text of the allegation. There was no further context for each case. The allegation text was heavily redacted by removing person names, organisation names, dates, etc., which resulted in the redacted text lacking grammatical correctness. With these limitations, the model did not outperform the baseline.

The Texas Board of Nursing (TBON) provided 3,000 cases with only the text of the allegation. No additional context was provided. TBON's text redaction was similar to that of the NMC, retaining grammatical structure. The system yielded a classification accuracy of 78% (9% above baseline).

Insert Table 1 here Comparing data and performance of three nurse jurisdictions

Table 1 Comparing data and performance of three nurse jurisdictions

Jurisdiction	Number of cases	Textual Data	Non-textual Data	Redaction	Baseline %	Accuracy %
NMC	1,250	Yes	Yes	Replace	62	71
AHPRA	1,300	Yes	No	Remove	65	65
TBON	3,000	Yes	No	Replace	69	78

Expert Testing

Between March to May 2021, case managers in each jurisdiction were invited to test the live tool and provide feedback on its utility and usability through an online survey and live discussion. In total, 22 case managers took part. Each case manager was given 4 cases to review. For each case they were requested to compare and annotate the similarity of the case to previous cases, review the relevance of each category, and make their own judgement of the risk along with their reasoning. The online survey required users to provide anonymised feedback on the tool in terms of usability, usefulness, response time, quality of risk predictions, case comparisons and comparisons with the relevant regulatory code or standard, and comment on additional functions. Qualitative feedback on the utility and usability of the tool was positive. A full analysis of the technical outputs from this testing stage is being reported in a separate publication.

Focus Groups on ethical implications

During February and March 2021, the Research team met with a group of regulatory experts in each jurisdiction. The purpose of the focus group was to seek views on the ethical implications of AI, and specifically the perceived barriers and benefits of machine learning tools in regulatory environments. Participants completed a consent

form, and the discussion was recorded and transcribed verbatim. Qualitative analysis using NViVo was carried out, generating a thematic analysis of the data. The consensus was that the tool had the potential to deliver consistency in decision making and efficiencies in working practices, improve transparency and provide both educational and training opportunities which could ultimately lead to improved defensible decision making. A full analysis of the outputs will be reported in future work.

Discussion

Our goal was to establish whether or not machine learning tools could be applied to the early stage of complaints handling in ways that maximized timeliness and accuracy and adhered to the principles of transparency and accountability that are fundamental to good regulatory practice worldwide. The results suggest that such tools are possible. The tool achieved good levels of accuracy in predicting the risk of the allegation by using NLP, combined with other non-textual features that provide context to the allegation. However, it also identified the need for more data to increase the levels of accuracy required to incorporate the tool into day-to-day decision making due to limitations in data provided. Our experiments (Table 1) showed that our tool yields considerably better performance when it has access to more data (either in the form of texts or categorical features) and more details of the complaints. This highlighted the importance of collecting more data, calling for more regulatory bodies in joining the development of such tools by sharing more data. Were such tools to be incorporated into routine use by regulators rather than a small subset of data, their potential would increase further. Critically the tool is able to identify the presence of harm in a given allegation, using textual and non-textual features (when available). This aligns closely with existing goals in nurse regulatory bodies and elevates the need to consider the context (for example previous history, access to supervision) of a complaint in regulatory decision making (NMC 2021).

Much of the literature on AI focuses on the need to apply clear and consistent principles of transparency in the design, testing, implementation and ongoing revision of any new tool. In this project, regulatory experts, case managers and in-house data scientists in each jurisdiction were involved in every stage of the development of the tool. The final prototype allows case managers to see how the tool has arrived at its decision, highlighting key words and sentences responsible for the prediction of a given risk category. Case managers in turn can use these features to evidence their decision making. The tool also has the potential to add a layer of quality assurance on bias to

human judgements, making use of more data (past cases of a similar nature, regulatory rules and guidance) in arriving at a case by case decision. Furthermore, the researchers allowed the regulatory data scientists access to the tool's software so they had access to the methods that were used to make the decisions.

Conclusions

Regulators, like other government agencies whose purpose is to serve the public interest, are managing ever increasing amounts of data. In disciplinary work, the capacity to store more data together with the desire to understand more about the context of an allegation impacts on the processes used to arrive at regulatory decisions. AI tools offer the possibility of enhanced quality assurance through faster and potentially more accurate decision making. We conclude that the application of such tools aligns with the principles of right touch regulation (Professional Standards Authority, 2015) and risk-based approaches (International Council of Nurses, 1998; Benton et. al., 2019) in that they offer new ways to deliver regulation proportionate to risk. At a time when regulators are becoming keenly aware of the unsustainability of current disciplinary systems, this may well be a welcome innovation.

Ethical Approval

The Research Team and Royal Holloway University's legal counsel worked closely with each regulator to ensure that the necessary ethical approval, permissions, data impact assessments, information sharing agreements and legal agreements were in place before any data was shared.

Acknowledgements

This research was funded by the Centre for Regulatory Excellence, NCSBN. The authors would like to thank colleagues at the Texas Board of Nursing, the Nursing and Midwifery Council UK and the Australian Health Practitioner Regulation Agency who made this project possible.

References

Acemoglue, D., Restrepo, P. (2020). Unpacking skill bias: automation and new tasks. *American Economic Association*,110 p.256-61. https://www.aeaweb.org/articles?id=10.1257/pandp.20201063

Australian Health Practitioner Regulatory Agency (AHPRA) (2021). Regulatory Principles of the National Scheme. https://www.ahpra.gov.au/News/2021-03-18-Regulatory-principles-consultation.aspx

Algorithmic Accountability Act of 2019. Congress.gov, Library of Congress, 11 April (2019). https://www.congress.gov/bill/116th-congress/house-bill/2231

Alpadin, E (2016). *Machine Learning: the new Artificial Intelligence*. MIT Press.

Assale, M., Dui, L. G., Cina, A., Seveso, A., Cabitza, F. (2019). The revival of the notes field: leveraging the unstructured content in electronic health records. *Frontiers in Medicine,* 6.

Babuta, A, Oswald, M, Rinik, C. (2018). Machine Learning Algorithms and Police Decision making. Legal, ethical and regulatory challenges. *RUSI Whitehall Report* 3-18. https://rusi.org/publication/whitehall-reports/machine-learning-algorithms-and-police-decision-making-legal-ethical

Benton, D., Cleghorn, J., Coghlan, A., Reed, C., Rodriguez, A., Voyt, T. (2019). Acting in the public interest; learnings and commentary on the occupational licensure literature. *Journal of Nursing Regulation Supplement, Vol 10 (2) S1-40.*

Benton, D., Scheidt, L., Guerrero, A. (2020). Regulating disruptive innovation. Oxymoron or essential innovation? *Journal of Nursing Regulation*, 11(1), 24-28.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer Verlag.

Brooks, A. (2019). The benefits of AI; six societal advantages of automation. Rasmussen University. https://www.rasmussen.edu/degrees/technology/blog/benefits-of-ai/

Cam, A., Chui, M., Hall. (2019). Global AI Survey: AI proves its worth but few scale impact. https://www.mckinsey.com/featured-insights/artificial-intelligence/global-ai-survey-ai-proves-its-worth-but-few-scale-impact#

Cossins, D. (2018). Discriminating algorithms: 5 times AI showed prejudice. https://www.americanprogress.org/issues/immigration/news/2019/09/05/474177/know-daca-recipients-united-states/?wpisrc=nl_health202&wpmm=1

Cronquist D. (2013). Management of Complaints Against Nurses: Review, Investigation, Action, Resolution. *Journal of Nursing Regulation 4(1), p 25-33.*

Davenport, T., Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal, 6(2), p 94–98*. https://doi.org/10.7861/futurehosp.6-2-94

Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: *Human Language Technologies*, 1, p. 4171-4186.

Forbes, K (2020). Lessons from Australia's Robo debt debacle. https://mc.ai/lessons-from-australias-robodebt%E2%80%8B-debacle/

International Council of Nurses (1998). ICN on regulation; towards 21st century models. International Council of Nurses, Geneva, 1998.

Kent, J. (2019). Could Artificial Intelligence Do More Harm Than Good in Healthcare? https://healthitanalytics.com/news/could-artificial-intelligence-do-more-harm-than-good-in-healthcare

Leibon, G., Livermore, M. A., Harder, R., Riddell, A., Rockmore, D. (2016). *Bending the Law* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2740136

Levin, S. (2019). 'Bias deep inside the code': the problem with AI 'ethics' in Silicon

Valley. https://www.theguardian.com/technology/2019/mar/28/big-tech-ai-ethics boards-prejudice

General Medical Council (GMC) (2017). UK health regulator comparative data report 2016. https://www.gmc-uk.org/-/media/documents/uk-health-regulator-comparative-report-final-220217_pdf-73538031.pdf

Gianfrancesco, M. A., Tamang, S., Tazdany, J., Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine, 178(11), p 1544-1547*. https://doi.org/10.1001/jamainternmed.2018.3763

McKinney, S., Seiniek, M., Shetty, S. (2020). International evaluation of a AI system for breast cancer screening. Nature 577, 89-94. https://www.nature.com/articles/s41586-019-1799-6

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B. (2019). Interpretable machine learning: definitions, methods, and applications. https://arxiv.org/abs/1901.04592

National Council of State Boards of Nursing (NCSBN) (2020). Environmental scan A Portrait of Nursing and Healthcare in 20202 and Beyond. *Journal of Nurse Regulation 10 (4) Supplement 1, S1-S35.*

Nursing and Midwifery Council (2021). Understanding fitness to practice, taking account of context. https://www.nmc.org.uk/ftp-library/understanding-fitness-to-practise/taking-account-of-context/commitment-6/

Nursing and Midwifery Council (2019). Annual Report. https://www.nmc.org.uk/globalassets/sitedocuments/annual_reports_and_accounts/ftpannualreports/nmc-fitness-to-practise-report-2019-singles-linked-contents.pdf

Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations.

https://science.sciencemag.org/content/366/6464/447

Park, J. H., Shin, J., Fung, P. (2018). Reducing Gender Bias in Abusive Language Detection. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 2799-2804).

Polson, N, Scott, J (2019). *AIQ: How artificial intelligence works and how we can harness its power for a better world*. Black Swan Publishers.

Reimers, N., Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) p. 3973-3983.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the problem of human control* .Penguin Books.

Salton, G., Buckley, C. (1988). Term-Weighting Approaches in Automatic Text Retrieval.. *Information. Process. Management.,* 24, 513-523.

Shalev-Shwartz, S., Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

Sharkey, N. (2019). AI expert calls for end to UK use of 'racially biased' algorithms https://www.theguardian.com/technology/2019/dec/12/ai-end-uk-use-racially-biased-algorithms-noel-sharkey?CMP=Share_iOSApp_Other

Spittal, M., Bismark, M., Studdert, D. (2019). Identification of practitioners at high risk of complaints to health profession regulators*. BMC Health Services Research*, 19, 380. https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-019-4214-y#citeas

Şulea, O. M., Zampieri, M., Malmasi, S., Vela, M., Dinu, L. P., van Genabith, J. (2017). Exploring the Use of Text Classification in the Legal Domain. ASAIL, 2017 London, UK.
http://ceur-ws.org/Vol-2143/paper5.pdf

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Wang, W. Y. (2019). Mitigating Gender Bias in Natural Language Processing: Literature Review. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 1630-1640).

Susskind, D. (2020). *A World without Work* Penguin Random House.

Tata, S., Patel, J. M. (2007). Estimating the selectivity of tf-idf based cosine similarity predicates. *ACM Sigmod Record*, 36(2), 7-12.

Vayena, E., Blasimme, A., Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine*, 15(11), e1002689.

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, 5(2), 241-259.

Waltl, B., Landthaler, J., Scepankova, E., Matthes, F., Geiger, T., Stocker, C., Schneider, C. (2017). Automated extraction of semantic information from german legal documents. In IRIS: Internationales Rechtsinformatik Symposium. Association for Computational Linguistics.

Williams, A., Nangia, N., Bowman, S. (2018). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: *Human Language Technologies*, 1, p.1112-1122.

Woodford, I. (2020). the rise of MeTooBots: scientists develop AI to detect harassment in emails. The Guardian, 3 January 2020.
https://www.theguardian.com/technology/2020/jan/03/metoobots-scientists-develop-ai-detect-harassment?CMP=Share_iOSApp_Other

World Economic Forum (2020). The Future of Jobs.

https://www.weforum.org/press/2020/10/recession-and-automation-changes-our-future-of-work-but-there-are-jobs-coming-report-says-52c5162fce/

Zhao, J., Wang, T., Yatskar, M., Ordonez, V.,Chang, K. W. (2018). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: *Human Language Technologies*, 2,p. 15-20.