Use of Artificial Intelligence in Regulatory Decision-Making

Robert Jago, BA, M.Phil.(Cantab); Anna van der Gaag, MSc, PhD; Kostas Stathis, PhD; Ivan Petej, PhD; Piyawat Lertvittayakumjorn, MSc; Yamuna Krishnamurthy, MSc; Yang Gao, PhD; Juan Caceres Silva, PhD; Michelle Webster, BA, MSc, PhD; Ann Gallagher, SRN, RMN, BA, MA, PhD; and Zubin Austin, BScPhm, MBA, MSc, PhD

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. This article describes the tool and the overall process of its development. The AI tool was not designed to replace human judgment but rather to perform three data-driven decision support tasks: (a) an independent risk prediction of the case, (b) a comparison with previous similar cases, and (c) a cross reference to relevant parts of the regulatory standards or rules in each jurisdiction. Three nursing regulatory bodies 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 and supported the potential role of an AI-based tool such as this in improving the efficiency and effectiveness of decision-making in disciplinary processes in nursing regulation nationally and internationally.

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

Artificial intelligence (AI) tools are increasingly used to improve the quality and speed of processing large-scale data sets in commercial and public sector organizations worldwide (Cam et al., 2019; Susskind, 2020; Zhang et al., 2021). The COVID-19 pandemic has accelerated the use of AI (World Economic Forum, 2020), leading to increased levels of automation and the demand for new technical skills in the workforce. It has been argued that AI has the potential to make humans more productive across many sectors if society takes a human-centric approach to technological advances (Acemoglu & Restrepo, 2020).

Increasing evidence shows that AI models can match human 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 judgment and machine-learning platforms may increase validity and fairness when compared with human judgment 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 variety of sectors; however, to date, they have rarely been tested in a health regulatory environment. For example, in Australia, Spittal et al. (2019) 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 on multiple sources of evidence and an in-depth understanding of rules. The initial step of an allegation review is primarily a manual process that 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, who are responsible for reviewing the incoming complaints against nurses, is also likely to increase (Sanson, 2017). A report by the medical regulator in the United Kingdom on the activities of nine U.K. professional health regulators identified a 32% increase in complaints against health practitioners over the preceding 6 years (General Medical Council, 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 (Nursing and Midwifery Council [NMC], 2019).

In light of these facts, there is a clear and growing need for innovative tools to assist regulatory staff tasked with processing complaints, particularly in screening complaints for those that are low risk, in an effort to streamline regulatory processes. This project aimed to develop an AI-based tool for improving the consistency and efficiency of decision-making in the nursing complaints process with a primary focus on streamlining the screening

stage of investigation. Rather than replacing human judgment, the intent of the tool is to provide data-driven support to regulators that will facilitate consistent, efficient decision-making when reviewing complaints.

Background

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 & Ben-David, 2014). A 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 (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 & Kalakota, 2019; Brooks, 2019). The Environmental Scan Report from the National Council of State Boards of Nursing (2020) suggested that in areas such as health screening and diagnostics, AI-enabled automated processes and AI-assisted patient engagement are growing rapidly, and they will have legal and ethical implications for regulators. While the use of AI technologies within healthcare to date shows 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, transparency, and the protection of human rights (Benton et al., 2020). For example, Gianfrancesco et al. (2018) showed that AI systems applied within clinical decision support can potentially exhibit important societal biases; if these systems are used incorrectly, they can amplify healthcare disparities. Obermeyer et al. (2019) reported that a machine learning algorithm used by many U.S. 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, reservations have been expressed (Ghosh & Kandasamy, 2020) and healthcare regulators have often been challenged by the perceived obscurity of the AI decision-making process (Davenport & Kalakota, 2019). Regarding accountability, Kent (2019) argued that because 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 ethical concerns, researchers have called for system-wide guidance, codes of practice, and even an ethical charter to ensure that AI development and use complies with ethical principles (Babuta et al., 2018; Council of Europe, 2018; McDonald, 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 (AI Asia Pacific Institute, 2020). In response, in the United States, legislators have implemented the Algorithmic Accountability Act (2019), which requires companies using high-risk automated decision support systems to conduct algorithmic impact assessments. In Europe, Article 22 of the General Data Protection Regulation states that "The data subject shall have the right not to be subject to a decision based solely on automated processing" and provides the data subject with an avenue for explanation and challenge.

These provisions have 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 interpretable algorithms, where the models provide insight about the inferences made about the data (Murdoch et al., 2019), and noninterpretable ("black box") algorithms, which digest large data sets without being able to demonstrate their workings (Babuta et al., 2018). In the case of black box algorithms, those from whom the data are derived have no knowledge of the decisions that have been made about them with the help of an algorithm (Babuta et al., 2018). Therefore, these algorithms should be subject to particularly high levels of testing and ongoing scrutiny (Vayena et. al, 2018). Babuta et al. (2018) call for system-wide guidance and codes of practice to ensure that AI development and deployment complies with ethical principles, including technical transparency and specifications about the availability of source code.

Considerations for an AI-Based Decision-Making Tool

In designing our system, we were aware that the human consequences of a complaint can be far reaching for the individual as well as their family and wider community. Our goal was therefore to design a system that was as accurate, transparent, unbi-

ased, and accountable as existing processes with the potential to improve these processes in terms of time, efficiency, and confidence. The following section describes the methods we used to attempt to realize this goal. We hope these methods will be helpful to researchers designing similar systems in the future.

Methods

We developed a web-based application that users could access via a password-protected portal. This was the most appropriate design because it 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. Initially, we used open access complaints data from the financial sector as part of our preliminary modeling work. This initial step allowed testing of various combinations of approaches before using nurse complaints data to develop the prototype.

Figures 1 through 3 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 1 depicts a table of all the uploaded cases and charts summarizing the statistics of the predictions. By clicking a row in the table, users are redirected to a results page (Figures 2 and 3) for the specific case, showing outputs of the system including the predicted risk score, the probability and confidence calculation that this score is correct, the key words used in calculating the risk score, and, for comparison, examples of similar cases relevant to the current decision and the regulatory rules pertinent to the case. In addition, users can provide feedback in response to the system outputs by 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.

Three nursing regulatory bodies—the NMC of the United Kingdom, the Australian Health Practitioner Regulation Agency (AHPRA), and the Texas State Board of Nursing (TBON)—agreed to participate in the research. 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 from the 5,700 complaints from the three jurisdictions were shared. All three regulatory bodies shared the same aspiration to explore data-driven solutions to the challenges of processing high volumes of complaints.

Risk Prediction

The tool classifies the complaint as high risk or low risk and provides the probability of the risk prediction. This prediction is achieved by using a "supervised machine learning" approach, where prominent characteristics of cases that constituted high risk or low risk to the public are learned from past cases that were processed manually. Such data are referred to as "training data," and they allowed us to begin to build a tool that could predict decisions about complaints using mathematical calculations of risk levels and previous judgments by humans with speed and accuracy.

We used a technique called "ensemble learning" (Wolpert, 1992) to create several machine-learning models (i.e., base models) and combined their predictions to provide the final output (Lertvittayakumjorn et al., 2021). The advantage of this technique is that it combines complementary strengths of the base models to enhance the system accuracy. This technique 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), which involves changing gender text in the prototype (e.g., "he" to "she"). 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 that 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 (Figure 2).

Similar Cases Retrieval

We hypothesized that it would be helpful to link past complaints that were semantically similar to the new complaint so as to help case managers cross-check with previous judgments and improve consistency in decision-making. First, we used a word-level similarity technique called "term frequency-inverse document frequency" to convert words into numerical vectors (Salton, 1988) in order to shortlist past complaints based on the overlapping of prominent words (Tata & Patel, 2007). Then, we fine-tuned the deep learning model Bidirectional Encoder Representations from Transformers, a powerful natural language processing model developed on a large volume of text data (Devlin et al., 2019), and we 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 regulatory body 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 & Gurevych, 2019) and textual inference (Williams et al., 2018), which related parts of the complaint to rules. This process identified the three most relevant rules to the case under consideration (Figure 3).

User Feedback

To improve the system over time, we collected and used feedback from the users throughout the development of the tool to design

FIGURE 1

Illustration of Multiple Cases, Risk Chart Summary, and Risk-Level Predictions

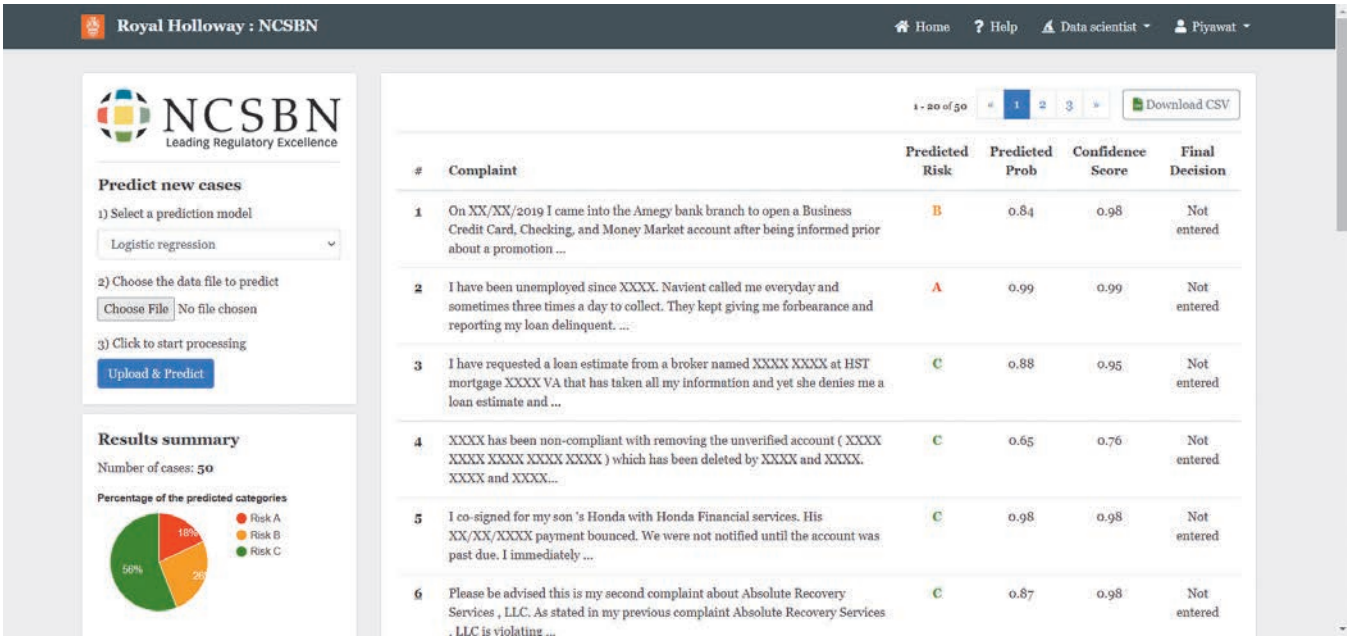
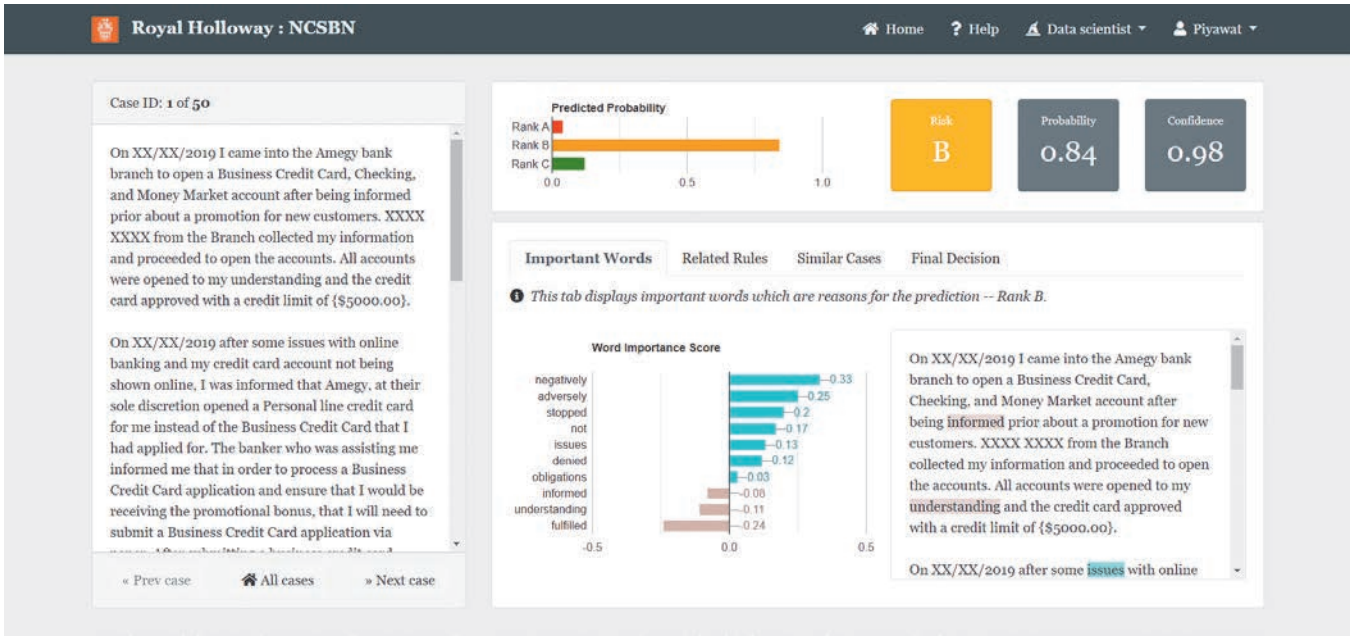


FIGURE 2

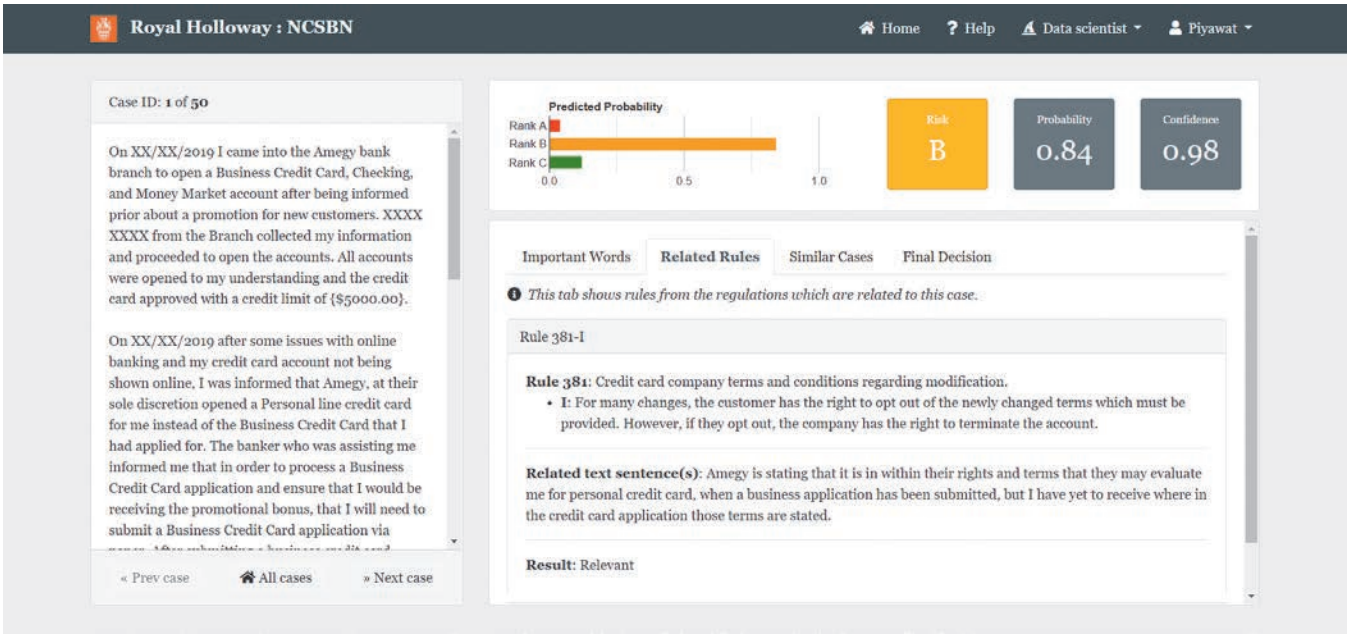
Illustration of a Single Case With the Risk Level and Ratings of Word Importance Scores



and fine tune the three features previously described. Case managers provided their own judgment 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 because there was insufficient training data to build supervised machine-learning models for them.

FIGURE 3

Illustration of a Single Case With the Relevant Regulatory Rules Shown



Results

Data Sets and Risk Prediction Accuracy

Table 1 provides an overview of the data and risk prediction accuracy for each jurisdiction. The NMC provided a data set of 1,250 cases. Each case contained the redacted text of the allegation and additional nontextual 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], [ORGANIZATION], [DATE], etc., so the resulting text was more or less grammatically correct and understandable. AHPRA provided a data set of 1,300 cases with only the text of the allegation—no further context was provided for each case. The allegation text was heavily redacted by removing person names, organization names, dates, etc., which resulted in the redacted text lacking grammatical correctness. The TBON provided a data set of 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, in which redacted text was replaced with tokens. All three data sets contained high- and low-risk cases. NMC and TBON data also provided medium risk cases, but they were small in number and therefore were included in the high-risk category for the purposes of testing the tool, as we were particularly interested in the tool's ability to discriminate low-risk cases. We then used this data set to test the accuracy of risk predictions for new cases.

The aim of the tool at this stage of development was to predict whether each case was of low or high risk. The accuracy of the prediction was measured by the percentage of correct risk predictions for the test cases (cases for which we already have the regulatory body's rulings) when the model is run. We compared the tool's prediction accuracy against the baseline prediction accuracy. The baseline predictions are done without a model or tool. In a data set that contains different categories of data appearing with different frequencies, the baseline predicts the most frequent category for all inputs. For example, if the labelled data set has 300 high-risk cases and 700 low-risk cases, the baseline will always label new cases as low risk before the model is run. Since the NMC data had grammatical structure and additional nontextual information, the tool's prediction accuracy for NMC data was 71% (9% above the baseline of 62%). TBON data also had grammatical structure but no additional nontextual information. However, TBON provided twice the number of cases. Hence, the tool yielded a predication accuracy of 78% (9% above the baseline of 69%). The tool's prediction accuracy was lowest for APHRA and did not perform better than the baseline (65%), likely because the data did not include grammatical structures. These differences in outcome provided valuable learning for future work because they demonstrated the relative impact of grammatical structure, nontextual information, and size of the data set.

An examination of the small number of miscategorized cases revealed that there were examples of high-risk cases miscategorized as low risk and vice versa. However, for NMC, it was more likely that low-risk complaints were miscategorized as high risk, whereas for AHPRA and TBON, the situation was the opposite. These differences may have been the result of additional nontextual information or the different proportions of high-risk and low-risk cases.

TABLE 1

Use of Artificial Intelligence in Regulatory Decision-Making: Data and Performance Comparison of Three Nursing Jurisdictions

Jurisdiction	Number of Cases	Textual Data	Non-textual Data	Redaction	Baseline %	Accuracy %
United Kingdom (NMC)	1,250	Yes	Yes	Replace	62	71
Australia (AHPRA)	1,300	Yes	No	Remove	65	65
United States (TBON)	3,000	Yes	No	Replace	69	78

Note. NMC = Nursing and Midwifery Council; AHPRA = Australian Health Practitioner Regulation Agency; TBON = Texas Board of Nursing.

Another feature of the tool was its ability to de-bias data. Gender de-biasing was completed on all three data sets. However, full bias testing was only possible with NMC data as the other two data sets did not have sufficient gender information. The bias testing on NMC data showed that the model did not use gender in making its decisions, possibly due to the redacted nature of the text. The gender swapping technique can further reduce the bias in the tool while sacrificing the accuracy by less than 1%.

Expert Testing

Between March and 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 four cases to review. For each case, the case manager was asked to compare and annotate the similarity of the case to previous cases, review the relevance of rules or standards identified by the tool, and make their own judgment of the risk along with their reasoning. The online survey required users to provide anonymized 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, as well as to comment on additional functions. Qualitative feedback on the utility and usability of the tool was positive. A full analysis of all the technical aspects of the testing stage is reported in a separate article (Lertvittayakumjorn et al, in press).

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 that could ultimately lead to improved defensible decision-making. It was noted that ensuring nondiscrimination through the safeguarding of access to and reinforcing the protective nature of a complaints process should never be compromised using such a tool. A full analysis of the outputs is planned in a forthcoming article (Austin et al, 2021).

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 prototype tests in each jurisdiction suggest that such tools are possible.

However, testing also identified the need for more cases to increase the levels of accuracy required to incorporate the tool into day-to-day decision-making. The tool achieved good levels of speed and accuracy in predicting the risk of the allegation by using natural language processing combined with other nontextual features that provide context to the allegation and large enough data sets, particularly for TBON and NMC. The AHPRA data yielded lower levels of accuracy on account of the lower levels of nontextual features and grammatical structures. However, once the prototype was delivered to them at the end of the project, AHPRA's data science team began undertaking further in-house testing, with higher levels of contextual features included.

Our experiments (Table 1) showed that the tool yields considerably better performance when it has access to more data (either in the form of text or categorical features) and more details of the complaints. This finding highlighted the importance of collecting more data, calling for more regulatory bodies in joining the development of such tools by ensuring the collection of similar data and the sharing of these data in an international repository. 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 nontextual features (when available). This process aligns closely with existing goals in nurse regulatory bodies and elevates the need to consider the context (e.g., 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 arrived at its decision, highlighting keywords 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 judgments, making use of more data (past cases of a similar nature, regulatory rules, and guidance) in arriving at a case-by-case decision.

Conclusion

This project highlighted the potential role and value of using AI-based tools to enhance efficiency and effectiveness of decision-making in nursing regulation. The success of this tool is based on the ability of AI to better manage and support analysis of large (and growing) data sets from diverse sources. For complaints and disciplinary processes, regulators must triangulate incomplete data from case files, precedents/similar cases, and regulatory standards/rules within an evolving context of public expectations regarding accountability and procedural fairness. We have demonstrated that AI tools offer the possibility of more methodical and systematic data management to support human decision-making and to facilitate enhanced quality assurance and quality improvement practices. Ultimately, this may improve both the quality of decision-making as well as the efficiency of regulatory processes. Our hope is that other regulators will replicate this work and build on it within the health complaints process as well as within other regulatory functions such as registration and accreditation, reducing costs and improving efficiency without compromising quality in 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 (Styles et al., 1997; 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.

References

Acemoglu, D., & Restrepo, P. (2020). Unpacking skill bias: Automation and new tasks. *American Economic Association*, 110, 356–361. https://doi.org/10.1257/pandp.20201063

AI Asia Pacific Institute. (2020). Lessons from Australia's Robodebt debacle. https://aiasiapacific.org/2020/07/17/lessons-from-australias-robodebt-debacle/).

Algorithmic Accountability Act of 2019, H.R. 2231, 116th Cong. (2019). https://www.congress.gov/bill/116th-congress/house-bill/2231

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,* 66.

Austin, Z., Jago, R., van der Gaag., A., Webster, M., Gallagher, A., Lertvittayakumjorn, P., Petej, I., Gao, Y., Krishnamurthy, Y., & Stathis, K. (2021). *Artificial intelligence in health professions regulation: qualitative results of an exploratory study in nursing* [Unpublished manuscript]. Department of Law and Criminology, Royal Holloway, University of London.

Babuta, A., Oswald, M., & Rinik, C. (2018, September). Machine learning algorithms and police decision-making: Legal, ethical and regulatory challenges [Whitehall Report 3-18]. Royal United Services Institute for Defence and Security Studies. https://static.rusi.org/201809_whr_3-18_machine_learning_algorithms.pdf.pdf

Benton, D. C., Cleghorn, J., Coghlan, A., Damgaard, G., Doumit, M. A. A., George, J. L., González-Juarado, M. A., Ewek, P.-E., Johansen, C., Msibi, G. S., Nyante, F., Owyer, E., Reed, C. M., Rodriguez, A., & Voyt, T. (2019). Acting in the public interest: Learnings and commentary on the occupational licensure literature. *Journal of Nursing Regulation, 10*(2 Suppl), S1–S40. https://doi.org/10.1016/S2155-8256(19)30120-6

Benton, D. C., Scheidt, L., & Guerrero, A. (2020). Regulating disruptive innovation: Oxymoron or essential innovation? *Journal of Nursing Regulation, 11*(1), 24–28. https://doi.org/10.1016/S2155-8256(20)30057-0

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

Brooks, A. (2019, November 4). The benefits of AI: 6 societal advantages of automation. Rasmussen University. https://www.rasmussen.edu/degrees/technology/blog/benefits-of-ai/

Cam, A., Chui, M., & Hall, B. (2019, November 22). 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, April 12). Discriminating algorithms: 5 times AI showed prejudice. *NewScientist.* https://www.americanprogress.org/issues/immigration/news/2019/09/05/474177/know-daca-recipients-united-states/?wpisrc=nl_health202&wpmm=1

Council of Europe. (2018). *CEPEJ European Ethical Charter on the use of artificial intelligence (AI) in judicial systems and their environment*. https://www.coe.int/en/web/cepej/cepej-european-ethical-charter-on-the-use-of-artificial-intelligence-ai-in-judicial-systems-and-their-environment

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

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). 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* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

General Medical Council. (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

Ghosh, A., & Kandasamy, D. (2020). Interpretable artificial intelligence: Why and when. *American Journal of Roentgenology, 214*(5), 1137–1138. https://doi.org/ 10.2214/AJR.19.22145

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

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, April 30). *Bending the Law.* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2740136

Lertvittayakumjorn, P., Petej, I., Gao, Y., Krishnamurthy, Y., Van Der Gaag, A., Jago, R., & Stathis, K. (2021). Supporting complaints investigation for nursing and midwifery regulatory agencies. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations* (pp. 81-91).

Levin, S. (2019). 'Bias deep inside the code': The problem with AI 'ethics' in Silicon Valley. *The Guardian.* https://www.theguardian.com/technology/2019/mar/28/big-tech-ai-ethics-boards-prejudice

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

McKinney, S. M., Seiniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M. Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., … 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. *PNAS, 116*(44), 22071–22080. https://doi.org/10.1073/pnas.1900654116

National Council of State Boards of Nursing. (2020). NCSBN's environmental scan: A portrait of nursing and healthcare in 2020 and beyond. *Journal of Nursing Regulation, 10*(4 Suppl 1), S1–S35.

Nursing and Midwifery Council. (2019). *Annual Fitness to Practise Report, 2018–2019*. https://www.nmc.org.uk/globalassets/sitedocuments/annual_reports_and_accounts/ftpannualreports/nmc-fitness-to-practise-report-2019-singles-linked-contents.pdf

Nursing and Midwifery Council. (2021, March 29). *Understanding fitness to practice: Taking account of context*. https://www.nmc.org.uk/ftp-library/understanding-fitness-to-practise/taking-account-of-context/

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453. 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). https://doi.org/10.18653/v1/D18-1302

Professional Standards Authority. (2015). Right touch regulation revised. https://www.professionalstandards.org.uk/docs/default-source/publications/thought-paper/right-touch-regulation-2015.pdf?sfvrsn=eaf77f20_20

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* (pp. 3973–3983).

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513–523.

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

Spittal, M. J., Bismark, M. M., & Studdert, D. M. (2019). Identification of practitioners at high risk of complaints to health profession regulators. *BMC Health Services Research, 19*, 380. https://doi.org/10.1186/s12913-019-4214-y

Styles, M. M., Affara, F. A., & the International Council of Nurses. (1997). *ICN on regulation: Towards 21st century models*. International Council of Nurses.

Sulea, 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. In *Proceedings of the Second Workshop on Automated Semantic Analysis of Information in Legal Text*. 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. https://doi.org/10.1145/1328854.1328855

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

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. R. (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, Volume* 1 (pp. 1112–1122).

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

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, 5(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

World Economic Forum. (2020). *The future of jobs*. https://www.weforum.org/reports/the-future-of-jobs-report-2020

Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., Lyons, T., Manyika, J., Niebles, J., Sellitto, M., Shoham, Y., Clark, J., & Perrault, R. (2021). *Artificial Intelligence Index Report 2021*. Stanford University. https://aiindex.stanford.edu/report/

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, Volume 2* (pp. 15–20).

Robert Jago, BA, M.Phil. (Cantab), is the Head of Department, Law and Criminology, Royal Holloway University of London, England. **Anna van der Gaag, MSc, PhD,** is a Senior Research Fellow, Royal Holloway University of London and Visiting Professor, Ethics and Regulation, University of Surrey, England. **Kostas Stathis, PhD,** is a Professor of Computer Science, Royal Holloway University of London. **Ivan Petej, PhD,** is a Research Fellow, Royal Holloway University of London. **Piyawat Lertvittayakumjorn, MSc,** is a Research Fellow, Royal Holloway

University of London. **Yamuna Krishnamurthy, MSc,** is a Research Fellow, Royal Holloway University of London. **Yang Gao, PhD,** is a Lecturer, Department of Computer Science, Royal Holloway University of London. **Juan Caceres Silva, PhD,** is a Research Fellow, Royal Holloway University of London. **Michelle Webster, BA, MSc, PhD,** is a Lecturer, Department of Social Science, Royal Holloway University of London. **Ann Gallagher, SRN, RMN, BA, MA, PhD,** is a Professor of Care Education, University of Exeter, England. **Zubin Austin, BScPhm, MBA, MSc, PhD,** is a Professor, University of Toronto, Ontario, Canada.

This study was funded by the National Council of State Boards of Nursing Center for Regulatory Excellence.

The authors would like to thank colleagues at the Texas Board of Nursing, the Nursing and Midwifery Council U.K., and the Australian Health Practitioner Regulation Agency who made this project possible.