Past Event: 2024 NCSBN Scientific Symposium - Advancements in Regulation: An investigation into the Potential of Artificial Intelligence to Support Regulatory Decision Making in Complaints about Nurses in the U.S., UK and Australia Video Transcript

Event
2024 NCSBN Scientific Symposium

More info: https://www.ncsbn.org/past-event/2024-ncsbn-scientific-symposium

Presenters
Anna van der Gaag, CBE, Visiting Professor University of Surrey
Robert Jago Professor of Law, Royal Holloway, University of London

- [Anna] Well, good afternoon, everyone. And I'm really delighted to be here, and honored to have the opportunity to actually present this research to you in person. I'm joined remotely by my colleague, Professor Rob Jago from Royal Holloway, University of London, who I hope is with us.

I think he is.

- [Prof. Jago] I am here, Anna. Thank you.

- Great. So Rob's going to be talking in a little bit about some of the qualitative findings from our work together. So the first thing to say is I'm not a technical person. So those of you who are computer scientists in the room, you're going to be disappointed because I'm going to talk very superficially about the actual technical side of this work.

I have a clinical background. I spent many months in the company with people with really impressive technical skills, and some of them are listed on this slide. And I'm sorry you can't see their names because they have been absolutely critical to the success of this project.

And I also want to thank Maryann in particular from NCSBN, but to the Center for Regulatory Excellence for taking a leap of faith back in 2018 when AI was hardly in the public discourse, and certainly not in relation to regulation, and for funding this project. So what this study was seeking to do was to explore the uses of a particular type of artificial intelligence in nurse regulation across three jurisdictions.

So here in the U.S., the UK, and Australia. And it was the first study of its kind in the world, as far as we're aware. And we were very fortunate to work with the Texas Board of Nursing who, again, were instrumental in making this happen. And I, again, want to acknowledge the contribution of Kathy Thomas, Mark Majek, who I know both retired now, but Dusty Johnson, Skylar Caddell, and Tony Diggs.

So I think Elise McDermott, I think is here from T-Bone [SP]. So will you please convey my heartfelt thanks to those people because they've been fantastic collaborators on this project? So I think when you think about artificial intelligence, you probably think about a whole range of different technologies.

And just yesterday, we heard about the robo call. Joe Biden's robo call for the New Hampshire elections. So, effectively, somebody trying to spread false information using the president's voice. That's the kind of thing that grabs the headlines.

That's the kind of technology that people are fearful of, and certainly don't want to see anywhere near their work environments. If you contrast that with the incredible successes of these kinds of technologies in, for example, improving the speed and accuracy of cancer diagnosis.

If you talk to some of the radiologists who are at the cutting edge of using AI in their clinical environments, they are incredibly excited about AI. So we've got a whole spectrum of interpretations, and constructs, and views about this technology in between those two extremes.

And I suppose one of the big challenges we face, whether it's in research, or clinical, or even indeed in commercial settings, is really whether we can actually understand AI as a force for good, whether it should be regulated, how it should be regulated, and where it will have most impact.

So I think these are the kinds of questions that are front of mind when people think about technologies. But there's one thing that's absolutely certain is that this technology will not go back in its box. It's here to stay. So in the context of regulation, our challenge was to try and explore how a particular type of artificial intelligence could be used.

And because Robert and I had been working in nurse disciplinary, the area of nurse discipline, and looking at the way complaints about nurses are handled, we were particularly excited about the prospect that these tools could actually help with the speed, and accuracy, and consistency of decision making in nurse discipline.

And what we know from research from around the world is actually that a very, very small number of nurses are a high risk to patients. And that actually, around about 70-plus percent of cases or complaints that are made to regulators around the world require no regulatory action.

So Robert and I were particularly interested in exploring how these tools could be used with that particular cohort where there was no harm to patients, and low risk in terms of regulatory assessment. And the other really important part of this, the thinking behind this work was that we were made aware of just how scarce a resource regulatory, disciplinary staff in regulation are.

So there's often a high turnover of regulatory staff who work in disciplinary teams because it's very high stress for relatively little reward.

And the number of complaints are increasing everywhere. So you've got this rise in the number of complaints. You've got a high turnover of regulatory staff. You've got a scarce resource that you want to look after. And if these tools could be shown to assist in that, then we would...that was essentially what we were aiming to do.

So our research question was, can this be done? Our aim was to focus first on whether or not an AI tool could actually calculate the risk level at the early stage of a complaint.

So when the summary data comes in, the complaint comes in, can a tool actually assess the data and come up with a risk classification in the way that we as humans do when we do this work? Secondly, we wanted to test whether the tool could actually link cases to regulatory rules and standards. And thirdly, could it detect, find previous similar cases in order for the person, the case manager who was making the decision about that particular case, and whether to take it forward could look at previous cases and see what the outcomes of the decisions had been, and add that to their judgment, their human judgment on the case that was in front of them?

In terms of our methodology, and as I said, I'm not going to go into great technical detail on this, we were fortunate to be able to access 3000 cases from Texas, and we had about the same proportion, 1200 or so from the UK and from Australia.

And we use that data to build the tool using a Python web development framework. And what this tool actually did was try a number of different AI classifiers. So we use five different AI classifiers.

For those of you who are from that background in the audience, we use gradient boosting, adaptive boosting, CNN, and an ensemble model, which was a combination of three of the five. And we fed in the complaint text, so all the information that we had about that complaint, including the source of the complaint, the risk level, and the harm to patients if there was any.

And then the classifiers would come up with a risk rating, or a risk classification on each case. That was the first task, if you like. And then the second, and we know from research and debate about AI that data quality is a big issue.

So what you put in to your machine learning tool, your training tool, determines the quality of information that you're going to get out of it. So we were keen to check, particularly for human bias on race, gender, and age. But at this early stage, we didn't have enough data on race and age, so we focused just on testing whether the tool could actually effectively a de-bias for gender.

So we used three different gender de-biasing techniques. So removing gender, neutral gender, and a gender swap so that we could make sure that the risk level that was calculated wasn't biased on gender.

And then the final part was the qualitative testing, which Robert's going to talk about shortly. We were very keen all the way through this project to involve regulatory staff, people on the ground who were doing this work, testing out what we were developing, involving them in the design, and getting their feedback.

So what I'm going to do now is just to show you some screenshots of the prototype that we've developed. For obvious reasons, I'm not using Texas Board of Nursing data, which wouldn't be appropriate. But this is showing accessible data from a U.S. financial regulator, in fact.

And as you can see, the first page is a secure login. What comes up when you go into the login page is, first of all, the summary. So you see there in the center, the cases... I'm so sorry, this is so small, but I know you're going to get these slides afterwards, so you'll be able to look at a bit more detail.

But if I use the pointer, this is just the headline from each case. You can see the numbers going down. And what you have going to the left of the slide is first the probability score, the confidence score. So that tells you something about how much confidence you can have that the risk level is actually accurate.

So this one says 98%. And then the final column, which is blank at the moment, is giving you the space to enter the human judgment. So right from the start, in the design of the dashboard, what you have is a tool that's leaving the final decision to the human.

So it's providing information, but it's not making a decision. That's left to the case manager. What you see in the bottom right corner is just the kind of graphics on a high medium and low risk. Again, that will be a calculation on each case.

This is perhaps a little bit clearer for those of you at the front of the room. So on my right, you see the actual text, the free text, if you like. And the sense of what you see is the key words that the tool has used to calculate the risk.

And it gives you a sense of which key words are of particular importance in arriving at that risk score. And then above, as you see, you've got your risk score, your probability, and your confidence score at the top of the page there.

So that's the risk calculation that the case manager can then use in forming their own judgment about the case. What we also developed, and this was very much in collaboration with regulatory staff who said, "These are things that we would find useful," is we, first of all, designed the tool so that it could show up the section of the regulatory rules.

And as we learned from working with Texas, there are pages and pages and pages of rules. So the tool effectively extracts the elements, the section of the rulebook that's relevant to that particular case.

So that's the second task, if you like, beyond the risk calculation. And the third is to allow the case manager to compare the case that they're looking at with any previous cases where a similar pattern of noncompliance has been found. So it's all about triangulation of data, and not in any sense about replacing human judgment.

And we've been very encouraged by the first phase of testing, the reliability testing here, where, as I said, we had these five different AI classifiers.

And we compared them to a baseline, and found that in terms of their reliability, the scores were looking promising. Not as high as we need them to be, but this is only on a sample of 1241 cases. So the first cases that we took through the phase one testing. And now I'm going to hand over to Robert.

I hope to talk to you about the qualitative findings.

- Okay. Thank you very much, Anna. And good afternoon, everyone. And thanks to the conference organizers for allowing me to appear remotely. So when we look at the general ethical concerns in the AI space, we found ourselves drawn to the seminal work of Michael Sandel. And you'll see on the slide that Sandel identified three key ethical concerns.

So the first of these are concerns as to privacy and surveillance. We then have the second, which is around bias and discrimination. And then the final ethical concern raised by Sandel is related to the role of human judgments, and the concerns about replacing human judgment. Now, in our work, we're also drawn, certainly in the area of health regulation, specifically to the work of Gabrielle Wolf.

And this has been critical because what they do is explore the ethical implications in the context of Australia, and health practitioner regulation. So here the first is dealing with equality before the law. Then you have transparency and accountability. The next concern is to do with consistency and predictability.

And then Wolf explores again the right to privacy as, of course, had been considered by Sandel. And then finally, there's reference to the use of AI, and how it could potentially undermine the right to work. Could you move to the next slide, please, Anna? So in our research, as well as developing and testing the prototype, we also, as Anna suggested, conducted focus groups with colleagues from our three sites.

And we explored with them what they saw as the benefits and burdens of using AI. And you'll see on the slide the three main themes that were raised here. Now, the first of these is negotiating trust and trustworthiness. So our participants focused much attention on trust, mistrust, and trustworthiness in relation to the inclusion of AI, and an AI tool in decision making related to complaints in nurse regulation.

Our focus group participants were very aware of the consequences of error in fitness to practice processes. And as they indicated, flawed decision making could result in registrants either continuing to harm patients, or, of course, being incorrectly judged to lack fitness to practice. Either way, the consequence is serious, and remind regulators of their responsibilities to ensure that decision-making processes are trustworthy.

Within this theme, our participants talked about prioritizing honesty and transparency. And they also said it was important to think about the language being used, and how impactful it could be in dealing with issues around AI. The second of our themes was about affirming fairness and nondiscrimination. And there was a real concern with this theme that any outputs from AI tools could potentially incorporate bias, and result in unfair and discriminatory decision making.

There was, therefore, a focus on trying to minimize bias. There was a need to avoid fabrication where it needed to be understood, which values or elements are focused in any decision-making tool or algorithm. And then also, related to this, was the final sub-theme, which was ensuring accountability.

And our participants explored the objective versus the subjective nature of decision making in this area, and the importance and challenges of clarity when it comes to accountability. Now, it was recognized that there would probably be some technical developments which assisted this process, but they should always remain alert to the potential for discrimination. And then the final of our themes was about managing burdens and benefits.

And there was a strong awareness in our focus groups that regulatory decision making is complex, and it needs to take into account context, uncertainty, and ambiguity. And therefore sub-themes here were talking about the shades of gray with the need for consistency, but obviously to negotiate complexity as well.

There were concerns of ensuring that nothing fell through the cracks. And there was a certain caution and optimism regarding what any AI decision support tool could actually deliver in the context of professional regulation. And it was critical, we thought, to mention that it was felt there was a real need for humility as to our expectations of AI.

And then there were some final discussions around effectiveness and burden reduction, and the potential benefit of any tool vis-a-vis the emotional content of fitness to practice complaints. Thank you. And I now hand back to Anna for some concluding thoughts.

- Thanks, Rob. So lots in there. And we really have packed in rather lot. We've published three papers, two in the JNR, and one in the "Journal of Computational Linguistics," which even I don't understand, but it's there. So if you're interested, please do read in more depth. And I think what I want to kind of leave you with, I suppose, is just that sense of how do we get from data to policy.

How do we do that in this AI-rich environment that we're increasingly living with? And I think we can't, as regulators, be left behind. We have to embrace this, but we have to do it in a way that reflects our values, and our commitment to fairness, to openness, to transparency, and to robust evidence.

So this study is a very first baby step towards that. And it demonstrates that you can use these multiple techniques of text classification, semantic similarity measurement, and natural language inference.

You can use these tools to actually provide an outcome which regulatory staff told us was of value to them. So they came into the work, a lot of them, with great skepticism about whether this was going to have any benefit for them and their work.

In fact, some of them were very honest about the fact that they were fearful that this was going to have an impact on them and their livelihoods, their jobs. And over the course of the two years of working with them, I think they could see, and they told us that they could see the benefits, as well as the potential pitfalls.

So we've deliberately designed this tool so that it can be used by anybody, any state board who's interested in replicating this study, who wants to continue with us on this journey. We have no copyright on the work.

And so we're really keen for NCSBN to take it forward. And I know from being here today that you have a fantastic team of data scientists who have the capacity. And I hope the interest and the willingness to take this work forward. It certainly isn't going to be me.

It's going to be the data scientists who do this. But just my very final thought is that transformation comes from people, and not from tools. And so we need to engage everybody in this debate because there is a lot of fear and anxiety about the intrusion of these sorts of tools into regulation, as in all areas of our lives.

And I hope that I've just given you, along with Robert, just a glimpse into the potential that this tool has to improve regulatory decision making, and to bring that consistency. And crucially, to shift that precious human resource to the high-risk cases, where we know there's going to be a full investigation, hugely time consuming.

Those high-risk cases, where there are patient safety implications are where we should be investing our human resource. And then embrace the new technology to make the screening out of the low-risk cases faster, better, more consistent without compromising that essential ingredient, which is human judgment.

Thanks very much. Robert and I would be very happy to take any questions, comments, objections, illuminations from you now.

- [Jose] Good afternoon. Sorry, that was a little too loud. Jose Castillo, Florida Board of Nursing. I love the [inaudible] comment. It's so enlightening that we are seeing AI being used in a very good light in the regulatory world.

As we all know, I'm also an educator. Not that you know. But as we all know in education, I should start there. AI is either being ostracized or it's being embraced completely. So it's like there's no middle ground. But I guess my question is from the regulatory realm, one of the aims is to calculate the risk level using anonymized data.

And I'm not sure, can you expand on that, like how much of the data? Because we know as regulators there's a plethora of data when it comes after investigation. So which ones would be highlighted by AI, or has that been looked at? How does that come to play with the overall evaluation, so that it will evaluate the risk?

- That's a very good question.

- Thank you.

- And quite possibly, my slide was a little misleading. So the aim in the initial stage, the very first pilot stage, was to see whether we could actually calculate risks using anonymized data. So that very first stage was about just seeing whether through the case summaries, we could come up with a risk score which was comparable to the human judgment without knowing anything more about the case.

As we went through the project, however, and clearly the Texas Board of Nursing Staff knew exactly the cases that they were giving us. They were the ones who were kind of familiar with the profiles, and the outcomes, and the context for these complaints. So as we moved through the project, we weren't using anonymized data, but we were de-identifying the cases, so as to protect people's identities.

So I probably slightly misled you on that. So the first stage was anonymized data from the financial regulator. That was a kind of database that our data scientists could gain access to. It's freely available.

But the next stage was actually a de-identification of cases, and testing to see whether that risk classification worked. So it was... Yeah, I hope that answers your question. Okay. Well, thank you so much for your attention.

Thank you.