How to Check the Reliability of Artificial Intelligence Solutions—Ensuring Client Expectations are Met

Jon Patrick[1]

[1] Health Language Analytics Global, Eveleigh, Australia

Appl Clin Inform 2019;10:269–271.

Address for correspondence Jon Patrick, PhD, MSc, BSc, Dip Land Surv, Grad Dip Behav. Health Psych, FACS, FACHI, MAMIA, Suite 4 International Business Centre, 2 Cornwallis Street, Eveleigh, 2015, Australia (e-mail: jon.patrick@hla-global.com).

Background and Significance

Artificial intelligence solutions for clinical tasks have been found to be prematurely released to clinical teams and thereby created increased risks and workload for clinicians. This letter discusses the issues that determine good AI practices.

A recent article in Forbes has described concerns in the United Kingdom over an artificial intelligence (AI) technology solution that diagnoses patient complaints and recommends the best course of action.[1] The article concentrates on the company Babylon but the critique is valuable for scrutinizing all AI products and their claims so as to offer herein generalizations of the issues and putative remedies that can be inferred from this case study.

The Forbes article says: "In the UK, Babylon Health has claimed its AI bot is as good at diagnosing as human doctors, but interviews with current and former Babylon staff and outside doctors reveal broad concerns that the company has rushed to deploy software that has not been carefully vetted, then exaggerated its effectiveness."

More broadly the Forbes article questions the relationship between tech startups and health organizations. Forbes says:

> "Concerns around Babylon's AI point to the difficulties that can arise when healthcare systems partner with tech startups. While Babylon has positioned itself as a healthcare company, it appears to have been run like a Silicon Valley startup. The focus was on building fast and getting things out the door…."

In particular, the gung-ho approach of information technology companies is identified: "Software is developed by iteration. Developers build an app and release it into the wild, testing it on various groups of live users and iterating as they go along."

A medical colleague has commented to me: "this is human experimentation and reckless."

Another commentary questioning Babylon's claims in more detail was published in the Lancet.[2] It made the assertions that: "In particular, data in the trials were entered by doctors, not the intended lay users, and no statistical significance testing was performed. Comparisons between the Babylon Diagnostic and Triage System and seven doctors were sensitive to outliers; poor performance of just one doctor skewed results in favor of the Babylon Diagnostic and Triage System. Qualitative assessment of diagnosis appropriateness made by three clinicians exhibited high levels of disagreement. Comparison to historical results from a study by Semigran and colleagues produced high scores for the Babylon Diagnostic and Triage System but was potentially biased by unblinded selection of a subset of 30 of 45 test cases."

> "Babylon's study does not offer convincing evidence that its Babylon Diagnostic and Triage System can perform better than doctors in any realistic situation, and there is a possibility that it might perform significantly worse… Further clinical evaluation is necessary to ensure confidence in patient safety."

Babylon has defended itself by saying that Babylon "goes through many, many rounds of clinicians rigorously testing the product … before deploying in the market." which appears somewhat contrary to the Lancet article, especially in the light of comments from a former employee: "Former staff say one of the biggest flaws in the way Babylon develops its software has been the lack of real-life clinical assessment and follow-up. Did people who used its chatbot ever go to an emergency room? If they did see a doctor, what was their diagnosis? 'There was no system in place to find out,' says a former staffer."

In a closing statement, Babylon answered criticism on the lack of publications of their work with the comment "The company admits it hasn't produced medical research," saying

received
January 23, 2019
accepted after revision
March 4, 2019

DOI https://doi.org/10.1055/s-0039-1685220.
ISSN 1869-0327.

it will publish in a medical journal "when Babylon produces medical research."

All of this reminds me of a panel I attended at Healthcare Information and Management Systems Society (HIMSS) in 2016 run by the IBM Watson team. They presented a set of comparative results based on diagnostic analyses from clinicians versus Watson where Watson proved the better in the match. Under close questioning from the audience, it turned out the clinicians were trainees and Watson was given credit for a correct answer if it had the answer in any of up to 50 possible diagnoses it offered. It seems they had never heard of false positives. The audience was unimpressed and Watson went on to generate a $60million burnout at MD Anderson Cancer Center with very little if anything to show for it.

What are the Problems?

So should we just condemn these behaviors or rather address the deeper issues raised here, that is: first, when should an AI solution be deployed in the field; how do you tell when it is ready and second, what are the consequences of prematurely releasing under-assessed pieces of AI technology.

The topics to be considered come under several headings:

- Public failure of any clinical AI solution casts AI technology for healthcare in a bad light and that is not good for those companies who are trying to solve seriously difficult problems with seriously complicated techniques.
- The Babylon technology is a chatbot, so it is using some form of natural language processing (NLP) or text mining to analyze patient content. The failures by Babylon pollute the clinical NLP field and slow down the adoption of other clinical NLP solutions.
- Premature deployment of an advanced technology is a bad idea as it produces dispiriting results and tarnishes everyone near and far.
- The notion that any AI company like Babylon cannot publish its results is specious and denies the importance of a transparent relationship between any clinical AI provider and the marketplace. Their work falls squarely in the field of health informatics and there are plenty of fora where its presentations would be welcome. But that would open them up to scrutiny from experts in their own field who would give a quite comprehensive and reliable assessment of their technology.
- Accuracy in any true AI-based NLP task above 90% is extremely difficult to attain and maintain, so sample collection, testing, and development methodologies need to be managed carefully and expertly.

What are the Remedies?

How can one best check the reliability of our AI solutions, whether the client is an intra-organizational party or an external paying client.

- Ensure there is an agreed and measurable accuracy rate for the client.
- Test very extensively and test to break. It is important to ensure a broad range of target content is collected for inclusion in the model building process and that flaws are actively sought and managed.
- Ensure there is a development methodology so that one can believe the accuracy rates the supplier calculates, that is, they are vigorously avoiding self-deception.
- Ensure the client is given weekly/fortnightly updates of their accuracy progress during the development process.
- Ensure the AI methods used have the ability to identify target content not present in the training data, that is, they are nondeterministic.
- Ensure the client is supplied with a mechanism to do independent testing.
- At go-live ensure they have an established methodology so that they continually monitor production results to detect false negatives which are then fed back into the training data so as to improve the computational model.
- The aim should be to do better than what is required by the client so that there is leeway if something crucial has been missed. Generally, a solution provider should not be comfortable if their internal testing is not producing a 4% greater accuracy than the agreed requirements. This is done to ensure that there is sufficient headroom to accommodate client document idiosyncrasies.

Some Tangible Examples

Example 1: Historical Records Coding

The California Cancer Registry has recently published its work[3] on a project to build a production line for coding histopathology reports to ICD O3. A processing pipeline was built from a corpus of 6,000 reports, using manually annotated texts for 150,000 phrases and an AI model was developed for those semantics that was correct to 99.7% at identifying those phrases and their lexicogrammatical variants.

Subsequently, a coding inference engine constructed for that corpus could correctly code to an accuracy of 97.4% for 140 body site locations and 133 histologies as well as grade, behavior, and laterality.

In the production usage of the system, the generalization and nondeterminism of the methods was validated by discovering 30% more body site locations and 80% more histology codes that were not in the original training data supplied by the client.

During the 15 months of the development and implementation, fortnightly reports on progress ran to 15 to 20 pages and presented the most recent accuracy results on all activities as well as other operational topics thus generating a closed-loop feedback system involving the provider and the client.

Example 2: Patient Outcome Predictor

In another project for Outcome Health (Melbourne, Australia) 57,000 general practitioner (GP) records were used to build an AI predictor for identifying patients with a high chance of being admitted to an emergency department based on information collected from GP visits.[4] The predictor was 74% correct for the 30 days following a GP visit by a patient. It was unreliable for any longer period of time. This was

important information and ensured that product claims were realistic and not overhyped. Ten GPs took part in the trial, of which nine found it valuable and its predictions credible and one did not.

The predictor is now being rolled out to over 400 GPs. The extensive testing of many different models in close collaboration with the client led to their confidence to instigate live trials, which in turn have led to a full rollout.

Incorporating a transparent feedback loop ensured that the client was intimately engaged in the design/development process and that the resulting solution was practical and worthwhile.

A Cautious Way Forward

Now we are all faced with understanding the level of risk that clinical NLP errors might trigger in patient care. A clinical colleague has quoted his earliest medical mentor, cardiothoracic surgery pioneer Victor Satinsky, with an important maxim that is equally applicable to prematurely released AI technology:

> "Critical thinking, always, or your patient's dead."
>
> Perhaps we could rewrite it for AI technologies as:
>
> Critical validation, always, or your client is worse off.

The end result is that by building trust with the right methodology and the client's close engagement, the supplier can deliver to client expectations. We can only hope that Babylon, Watson, and other health AI vendors will introduce more rigor into their methods in a transparent method so that healthcare can receive the best of what they can offer.

Pragmatic Checklist for Adopting a Clinical AI Approach

For improving the reliability of an AI solution:

- Ensure there is an agreed and relevant accuracy rate.
- Test very extensively using a closed-loop feedback process that involves the client.
- Ensure there is a development methodology that transparently calculates accuracy rates in a way acceptable to the client.
- Ensure there are weekly/fortnightly updates to the client of processing accuracy progress.
- Ensure that the AI methods used are nondeterministic.
- Ensure the client has a mechanism to do independent testing.
- Ensure there is an established methodology after go-live to continually monitor production results and adjust the AI models as needed.

Multiple Choice Questions

1. What is the accuracy threshold beyond which clinical NLP becomes very hard and costly to improve?
   a. 80%
   b. 85%
   c. 90%
   d. 95%

   Correct Answer: The correct answer is option c.

2. What do you do to assist a client in validating the AI you supply to them?
   a. Supply the document on how you built the technology.
   b. Supply weekly or fortnightly progress reports while building the technology.
   c. Ensure there is a development methodology that transparently calculates accuracy rates in a way acceptable to the client.
   d. Train the client to monitor the results of the technology when it is in production and report the result back to the provider.

   Correct Answer: The correct answer is option d.

Protection of Human and Animal Subjects

No human and animal subjects were involved in the project.

Funding

None.

Conflict of Interest

The author is the CEO of a company that was contracted to perform the two case studies described in this letter. All the facts quoted in the case studies are available from third-party sources for validation. Beyond that there is no conflict of interest.

References

1 Olson P. This health startup won big government deals—but inside, doctors flagged problems. Forbes 2018(December 17). Available at: https://www.forbes.com/sites/parmyolson/2018/12/17/this-health-startup-won-big-government-dealsbut-inside-doctors-flagged-problems/#2f3f47dbeabb. Accessed January 4, 2019
2 Fraser H, Coiera E, Wong D. Safety of patient-facing digital symptom checkers. Lancet 2018;392(10161):2263–2264
3 Moody C, Scocozza M, Brant M, et al. Using natural language processing to screen and classify pathology reports. NAACR Annual Conference June, 2017. Available at: https://www.naaccr.org/wp-content/uploads/2017/06/Using-Natural-Language-Processing-to-Screen-and-Classify-Pathology-Reports.pdf. Accessed January 4, 2019
4 Pearce CM, McLeod A, Patrick J, et al. POLAR diversion: using general practice data to calculate risk of emergency department-presentation at the time of consultation. Appl Clin Inform 2019; 10(01):151–157