Modelling the Impact of AI Systems for Clinical Decision Support

Mariana Neves and William Marsh

March 18, 2021

Modelling the Impact of AI Systems for Clinical Decision Support

Mariana Raniere[1, 2] and D. William R. Marsh[1, 3]

[1] 1 School of Electronic Engineering and Computer Science,
Queen Mary University of London, London UK
[2]m.r.neves@qmul.ac.uk, [3]d.w.r.marsh@qmul.ac.uk

Abstract. Many AI (or ML) systems have been proposed for clinical decision support, providing diagnosis, prognosis or treatment recommendation. It is well known that the impact of these systems varies, with benefits such as improved decision making or time saved set against any potential harm introduced by the AI. The main technique proposed to evaluate impact is an 'Impact Study', a form of trial of a completed system. Vital though such studies are, they require at least a prototype system to be deployed which can be expensive. Meanwhile, the merits of AI predictors are mainly argued using accuracy measures, such as a confusion matrix or an AUC. We argue that the impact of a proposed AI system should be modelled during its development, to justify the expense of an Impact Study. We show that an Influence Diagram can be used for this and provide a small set of models for generic AI systems, with two main findings. First, we show that the way that an AI predictor is used – primarily how it interacts with clinical decision makers – is at least as important as its predictive accuracy. Indeed, we show that different uses of the same predictor vary in impact without any change in its accuracy. Secondly, we show that the proposed use of an AI predictor also determines the information needed to model its potential impact. Some information is always needed on the decision accuracy of existing clinical decision makers, but the form and extent of this varies.

Keywords: Impact analysis, Clinical decision support, Influence diagram, Artificial Intelligence.

1 Introduction

1.1 Is AI for Decision Support Clinically Beneficial?

Systems based on AI (or ML) for clinical decision-support are developed to improve medical care. Such systems usually include a diagnostic or prognostic model built from patient data to diagnose diseases, predict the likely course of the disease or optimize the delivery of care. Although the terms used vary, it is clearly important to distinguish 'prediction' from the wider 'decision-support' system, into which the prediction goes.

It is well understood that the accuracy of the predictions is not sufficient to ensure that a system is clinically useful. The problem has been known at least since 1995 when

Wyatt and Altman [1] suggested that for a model to be clinically useful, it should have credibility, evidence of generality and accuracy and evidence of clinical effectiveness. This introduces a key distinction between accuracy (or validation) and effectiveness (or, more commonly, impact). Validation ([2], [3] and [4]) concerns evidence that the model is general, so that it is still accurate in a new context. To show this, prediction accuracy is evaluated with data from a place and time different from the original development.

In contrast, the impact of a decision support system concerns whether its use in a clinical setting benefits the delivery of care, for example by improving decision-making or reducing costs. A 2015 systematic review [5] found that 'relatively few' of the prediction rules developed have been evaluated in an impact study, although frameworks for these studies have been proposed ([4], [6], [7]). For example, in [7], a four-phased framework is proposed. The first phase is to check that the validation has been completed. The second 'preparation' phase assesses the acceptability of the decision-support system to its users and any potential barriers to the use (both organizational and individual). The third phase is an experimental study of changes to clinical care: several designs are possible including before and after or cluster randomized. If this is successful, long-term implementation follows.

The main argument of this paper is to suggest that two aspects of impact can and should be considered and used to estimate impact, long before the actual impact is measured in the 'experimental' phase of an impact study. These aspects are:

- The way the prediction system will be used for decision support, notably how it interacts with clinical decision making, and
- The hoped-for benefits, whether reductions in costs or workload or better decision-making.

The impact study frameworks do not neglect these issues: for example, in [7] an important aspect of the 'preparation' phase 'is determining how the *[prediction system]* will be integrated into the clinical workflow'. Similarly, it is suggested in [3] that potential impact should be considered (using simulated decisions) before conducting a study to assess actual impact. Further, [4] and [6] distinguish predictions designed to 'assist' (or support) decision-making from those that make decisions, with [3] pointing out that: "clinicians will not always follow the rule's recommendations: They may not consult the rule at all, they may apply it inaccurately or unreliably, they may deliberately overrule its recommendations, or they may be unable to implement its recommendations for various reasons".

We argue that the impact of a decision-support system should be estimated at an early stage. We present a method to do this and draw attention to the relationship between prediction accuracy, proposed used and impact. Accuracy does not determine impact independently of use. We also show that any estimate of impact requires some estimate of the performance of the existing delivery of care, but again the information needed vary with the proposed use of the AI system.

1.2 Outline

In section 2, we describe the modelling approach. Section 3 gives models for three differing uses. Our conclusions are in the final section.

2 Approach to Modelling Different Uses of AI system

2.1 How an AI System Can be Used Clinically

A prediction system can be 'integrated into the clinical workflow' in many different ways to form an AI-system clinical decision system. An important distinction is the role of the clinical decision-maker (a doctor or nurse for example) versus that of the AI system. At one extreme, the decision is fully automated, while at another the AI system and the clinician work together. Restricting our attention to diagnosis, **Table 1** shows three possible uses that we will analyze. These are intended to represent the much wider range of real possibilities.

Table 1. Diagnostic Uses of AI.

Name	Description of Use
Replace	The AI diagnoses all patients, replacing the clinician.
Filter	The AI diagnoses some patients; the clinician sees fewer.
Assist	The AI and the clinician work together.

The first two uses 'Replace' and 'Filter' are broadly directive (real examples include [8] and [9]), while the other is assistive. It is notable however, that many papers on predictors for clinical use say little about intended use. For example, in a recent collection of clinical AI systems, most gave only validation results [10].

2.2 Utility Models

The impact of introducing an AI system may include both benefit and harm. We can model this using a utility model. In section 3, we show such models for the different uses. Here we outline the parameters needed in the utility model and the potential areas of impact for each type of use. Our assumed parameters are intended only to be indicative; what matters is the forms of information needed to estimate impact and how it varies with use. We focus on two areas:

- Healthcare cost: we assume that the cost of a consultation with a clinician exceeds the cost of operating the AI; for simplicity, we assume the latter is zero. The cost of a clinical consultation is assumed to be 100 units, although additional costs may occur when the opinion of the clinician and the AI system differ. We have not considered treatment costs beyond the consultation. We also do not distinguish other related outcome, such as reduced workload or waiting lists, though this could be done.
- Patient outcome. We assume that the patient outcome depends on the presence of the disease and the diagnosis. In general, patient outcome is measured in quality-adjusted life-year (QALY). According to the National Institute for Health and Care Excellence (NICE), QALY is "a measure of the state of health of a person or group

in which the benefits, in terms of length of life, are adjusted to reflect the quality of life". This measure is used as a currency to compare healthcare interventions and it considers both the quantity and quality of life generated by them. The patient outcome we assume for each diagnosis and state of the disease is shown in **Table 2**.

Table 2. Example patient outcomes.

Disease	**False**		**True**	
Diagnosis	**False**	**True**	**False**	**True**
Patient Outcome	0	-50	-500	5000

The numbers shown in **Table 2** are indicative only: the consequences of the different errors (false positive and false negative) need to be assessed and can vary greatly in different circumstances. Our example gives the harm of a false negative a smaller magnitude than the benefit of a correct positive diagnosis; this might be appropriate when the result is a delay to the start of care. Similarly, we have even less harm from a false positive. A limitation on the use of utility models is that it might allow a (small) deterioration in the quality of care if compensated for by a sufficient cost reduction whereas this might not be permitted by a regulator; we note an example of this below. **Table 3** summarizes the potential areas of benefit for difference AI system uses.

Table 3. Summaries of Expected Benefit by Use.

Name	**Potential Expected Benefits**
Replace	Primarily cost saving, with equivalent patient outcome.
Filter	Cost reduction, by reducing the number of patients to be seen. Outcome may also change, however.
Assist	Primarily improved patient outcome; since all patients are still seen and the healthcare costs do not reduce (given our assumptions).

2.3 Performance Assumptions

In this section, we introduce some assumptions about the performance of both the AI system and clinician, as it will become clear that we cannot estimate the impact of an AI system without knowledge of the performance of an existing clinical care system.

Disease Prevalence: We assume disease is present in 30% of cases. With the error rates, the affects the number of errors of different types possibly impacting benefit.

Receiver Operating Curve (ROC) and Confusion Matrix. The performance of the AI system can be represented by its ROC, showing the tradeoff between sensitivity (or true positive rate – TPR) and specificity (or false positive rate – FPR). **Fig. 1** shows a simulated ROC. The false negative rate (FNR) can be calculated from the TPR and the true negative rate (TNR) from the FPR:

$$TPR = \frac{TP}{TP + FN} = 1 - FNR \qquad TNR = \frac{TN}{TN + FP} = 1 - FPR$$

Every point on the curve represents a possible ‘operating point’, from which a ‘confusion matrix’ can be calculated, for a known prevalence of the disease). The accuracy of decision-making can be summarized in a confusion matrix (see **Table 4**). We assume that the clinician’s confusion matrix is known, as shown for example in **Table 4**. In Section 4, we consider whether such an assumption is plausible.

Table 4. Confusion Matrices: in General, and for the Clinician (Doctor ‘B’, 30% prevalence)

		Actual	
		False	**True**
Pre-dicted	False	TN	FN
	True	FP	TP

		Actual	
		False	**True**
Pre-dicted	False	63%	9%
	True	7%	21%

Fig. 1 also shows 5 possible operating points P1-5 for the AI system, all lying on the curve. Two points are shown for the clinical decision maker, one (B) above the AI’s ROC and the other (A) below.

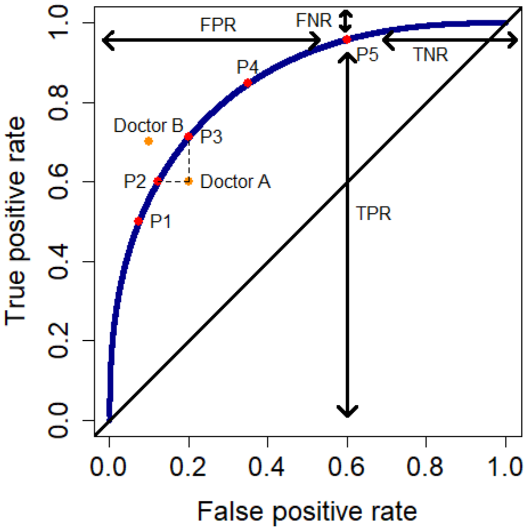

Fig. 1. Simulated ROC curve for the AI in which Area Under the Curve (AUC) is 84%.

3 Utility of an AI System in Different Uses

The utility calculations can be implemented as Hybrid Bayesian networks. Here we use the AgenaRisk toolset [11].

3.1 Modelling an AI System Replacing a Clinician

Fig. 2 shows a utility model (as an influence diagram) for an AI system replacing a clinician. The diagnosis depends on the use AI as does the healthcare cost. The patient outcome depends on the diagnosis and the presence of the disease. The performance of both AI (at the chosen operating point) and clinical decision makers are parameters in the model and determine the utility.

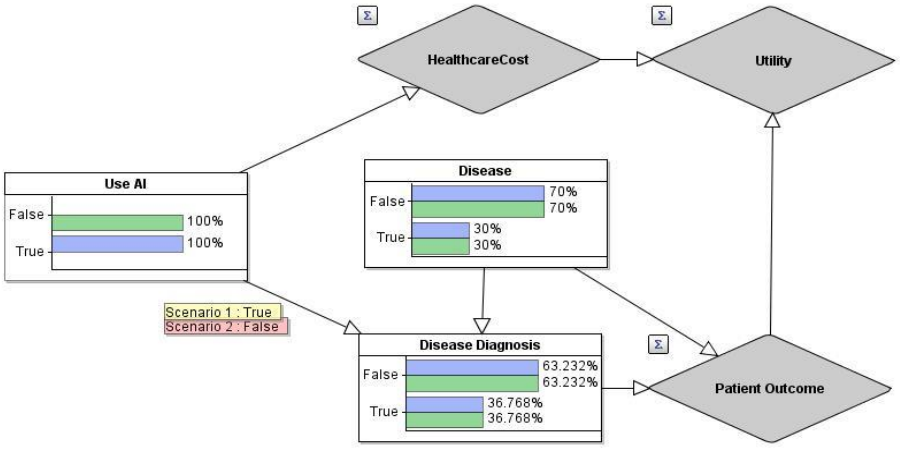

Fig. 2. Influence diagram for the case in which the clinician is replaced by the AI.

If the performance of the clinical decision-maker matches point A in **Fig. 1** then it is clear that the utility of using the AI system exceeds that of not using it, anywhere between point P2 and P3, as both the FNR and FPR are reduced. The clinician's utility (at point A) is 833.0; that of the AI systems at point P2 is 835.71 and at P3 is 1017.6. The use of AI use could be still beneficial (in terms of utility) if it made a few more mistakes than the existing decision maker, this may not be accepted in practice; instead it is more likely that the AI developers must show 'non-inferiority' of clinical outcome. The location of the best operating point depends on the ratio between the disutility of false negative and false positive diagnoses.

Summary of AI System Replacing a Clinician

The following table summarizes our results for this use of AI.

Table 5. Summary of Results for an AI System Replacing a Clinician.

Criteria	Finding
Positive benefit	AI performance must dominate the clinician's
AI performance	ROC curve
Clinical performance	Full confusion matrix

3.2 Modelling an AI System Filtering Patients Seen by a Clinician

In a diagnostic problem, an AI system can be used to filter out some cases so that a clinician sees fewer cases, saving clinical time (and money). We consider two cases a) no disease filter, where the AI only filters out the cases it is confident have no disease and b) disease present filter, the opposite. The model shown in **Fig. 3** is for case a): the variable 'Doctor Diagnosis Possible' represents the diagnosis given if the clinician were asked. The variable ‘Actual Diagnosis’ represents the final diagnosis: in case when ‘AI diagnosis’ is ‘false’ (i.e. no disease) then the Actual Diagnosis is false; if the AI response is “Don’t know”, then the doctor’s diagnosis is used as the Actual Diagnosis.

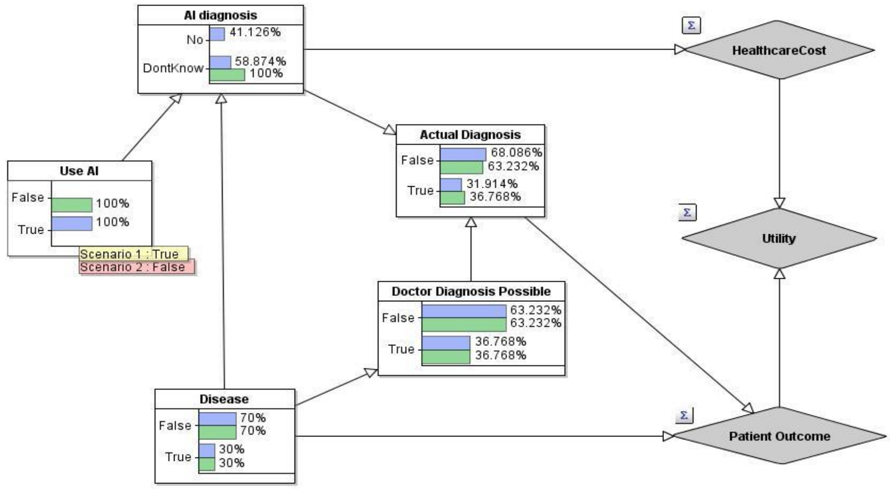

Fig. 3. Influence diagram for the case in which the AI filters patients.

Used in this way, it is possible for the AI system to be of benefit even if its performance does not dominate the clinician’s, since we can choose an operating point to minimize a certain type of errors. We have therefore compared to AI system to clinician B in **Fig. 1**– with better decision performance than assumed for the ‘replace’ case**.**

Table 6. Change in Utility when AI Filters Patients Seen by the Clinician.

Type of Filter	**AI System Operation Point**				
	P1	**P2**	**P3**	**P4**	**P5**
(a) No disease	-494.3	-385.5	-265.3	-124.9	-17.6
(b) Disease	225.1	266.6	310.7	358	384.3

It seemed ‘obvious’ to us that the ‘no disease’ filter would be beneficial, particular operated at P5 (to minimize FNR). The AI system simply detects some disease-free patients, leaving the rest to the clinician. The model does not support our intuition: the problem is that both AI and clinician have an FNR, so together miss more case than either operating alone. With the outcome utilities (and prevalence) we have assumed, the cost saving does not balance this. Instead, the ‘disease present’ filter has a benefit. The results are summarized in **Table 7.** .

Table 7. Summary of Results when AI Filters Patients Seen by the Clinician.

Criteria	**Finding**
Positive benefit	AI performance can be worse than the clinician’s.

	Benefit relative costs of FP and FN
AI performance	ROC
Clinician information	Confusion matrix (but one rate dominates).

3.3 Modelling an AI Used to Assist a Clinician

When an AI system is used to assist a clinician with a diagnostic decision, she has access to AI system's predictions before giving a final diagnosis to a patient. If there is a conflict, the clinician has a chance to re-evaluate her diagnosis.

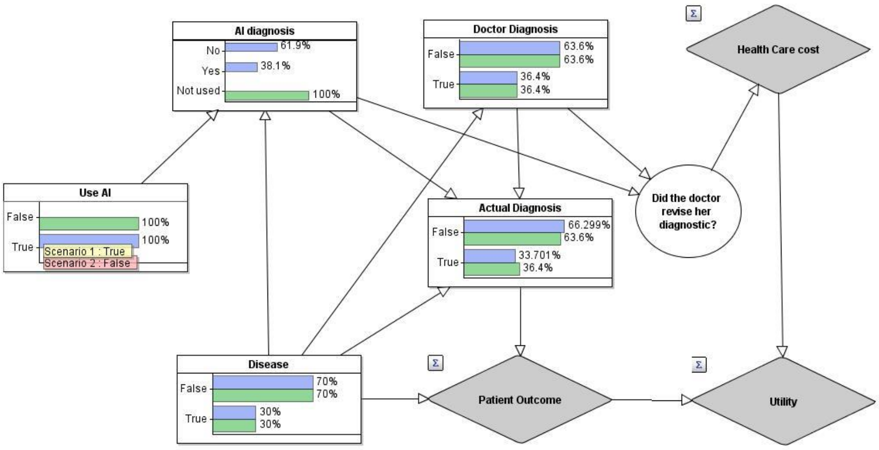

Fig. 4. Influence diagram for the case in which the AI assist the doctor.

A model of this situation requires some assumption about how the clinician and AI system interact. In the model of **Fig. 4**, we assume that when the doctor and AI disagree, the clinician revises her diagnosis in 50% of the cases in which she was wrong (but never when she was right). However, this has an associated cost, which arises whenever clinician and system disagree: this is assumed to be the same as the cost of the original consultation (100 monetary units) representing extra time and the costs of additional tests. Clearly, with these assumptions, costs do not reduce and any impact of using AI relates to a better patient outcome. We consider the plausibility of this model of the interaction and possible alternatives in the final section.

The following table shows the increased utility of using the AI at the different operating points P1 to P5:

Table 8. Increase in Utility when Clinician Assisted by AI.

Clinician	**AI System Operation Point**				
	P1	**P2**	**P3**	**P4**	**P5**
A	136.21	167.50	201.60	240.00	266.10
B	99.3	122.4	146.9	173.2	187.8

We note that a) the impact of this use of AI is generally smaller than the two earlier cases b) that it has greater impact for less accurate clinician A and c) the impact increases as the operating point of the AI has a lower FNR (eventually, the costs of the extra reviews exceed the benefits). The results are summarized in **Table 9**.

Table 9. Summary of Results for the Case where AI Assists a Clinician.

Criteria	Finding
Positive benefit	AI can be beneficial even if its performance is worse than the clinician's.
AI performance	ROC curve.
Clinician performance	Full confusion matrix. Model of interaction between AI and clinician.

Our assumption that the clinician is able to revise some percentage of incorrect decisions may be optimistic, since the cases that are difficult for a clinical decision maker may also be hard for an AI system.

4 Conclusions and Further Work

Our utility models have shown that the same AI system used in different ways can have a very different impact. In some cases, the performance of the AI system must exceed that of the clinical decision maker; in other cases, it can still be beneficial with lower performance. Making an estimate of impact requires the developer of an AI system for clinical use to consider both the potential benefits and the proposed use; we believe more systems would have impact of this was done.

We have also shown that, with the utilities we have used, the impact of assisting a clinical decision maker is less than replacing her. Moreover, estimating the impact of the 'assist' case requires much more information. In all cases we need some information about the performance of the clinical decision maker. Our assumption that this can be characterized by a simple confusion matrix is simplistic as, for example, different individuals may have different performance. We plan to investigate replacing fixed parameters here (and elsewhere) with distributions. A more profound issue is whether such information can be obtained. We argue that a full impact study would be pointless without, in the example of assistive AI, some indications that decisions need to be improved since there is no impact in assisting a perfect decision maker. Nevertheless, the calculation of confusion matrix requires a 'ground true' – a 'correct' outcome – for each decision and this may not be available. One approach might be to look at the consensus of a group of decision makers for similar patients.

A particular problem is that we know little about the interaction between a clinical decision maker and an AI system intended to assist with decision-making. We could consider the AI to behave like a 'second opinion' and the effects of this have been studied, for example in [12]. So far, we have assumed that mistakes are made independently, but there are many other effects to consider. For example, reducing the number of patients may change decision accuracy [13]. Furthermore, some decisions are harder than others and it is likely that those that the AI system gets wrong may also be incorrectly diagnosed by the clinician. We plan to investigate the impact of correlation between the decision performance of the two decision makers in future.

References

[1] J. C. Wyatt and D. G. Altman, “Commentary: Prognostic models: clinically useful or quickly forgotten?,” *BMJ*, vol. 311, no. 7019, pp. 1539–1541, Dec. 1995.

[2] D. B. Toll, K. J. M. Janssen, Y. Vergouwe, and K. G. M. Moons, “Validation, updating and impact of clinical prediction rules: A review,” *J. Clin. Epidemiol.*, vol. 61, no. 11, pp. 1085–1094, Nov. 2008.

[3] D. G. Altman, Y. Vergouwe, P. Royston, and K. G. M. Moons, “Prognosis and prognostic research: validating a prognostic model.,” *BMJ*, vol. 338, p. b605, May 2009.

[4] K. G. M. Moons, D. G. Altman, Y. Vergouwe, and P. Royston, “Prognosis and prognostic research: application and impact of prognostic models in clinical practice.,” *BMJ*, vol. 338, p. b606, Jun. 2009.

[5] E. Wallace *et al.*, “Impact analysis studies of clinical prediction rules relevant to primary care: a systematic review,” *BMJ Open*, vol. 6, no. 3, p. e009957, Mar. 2016.

[6] B. M. Reilly and A. T. Evans, “Translating clinical research into clinical practice: impact of using prediction rules to make decisions.,” *Ann. Intern. Med.*, vol. 144, no. 3, pp. 201–9, Feb. 2006.

[7] E. Wallace *et al.*, “Framework for the impact analysis and implementation of Clinical Prediction Rules (CPRs),” *BMC Med. Inform. Decis. Mak.*, vol. 11, no. 1, p. 62, Dec. 2011.

[8] “London hospitals to replace doctors and nurses with AI for some tasks | Society | The Guardian.” [Online]. Available: https://www.theguardian.com/society/2018/may/21/london-hospitals-to-replace-doctors-and-nurses-with-ai-for-some-tasks. [Accessed: 21-Jan-2019].

[9] S. Razzaki *et al.*, “A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis.”

[10] L. A. Celi, L. Citi, M. Ghassemi, and T. J. Pollard, “The PLOS ONE collection on machine learning in health and biomedicine: Towards open code and open data,” *PLoS One*, vol. 14, no. 1, p. e0210232, Jan. 2019.

[11] B. Yet, M. Neil, N. Fenton, A. Constantinou, and E. Dementiev, “An improved method for solving Hybrid Influence Diagrams,” *Int. J. Approx. Reason.*, vol. 95, pp. 93–112, Apr. 2018.

[12] J. G. Elmore *et al.*, “Evaluation of 12 strategies for obtaining second opinions to improve interpretation of breast histopathology: simulation study.,” *BMJ*, vol. 353, p. i3069, Jun. 2016.

[13] A. Wilson and S. Childs, “The relationship between consultation length, process and outcomes in general practice: a systematic review,” 1012.