Chapter 14
Trust and Transparency in Machine Learning-Based Clinical Decision Support

Cosima Gretton

Abstract Machine learning and other statistical pattern recognition techniques have the potential to improve diagnosis in medicine and reduce medical error. But technology can be both a solution to and a source of errors. Machine learning-based clinical decision support systems may cause new errors due to automation bias and automation complacency which arise from inappropriate trust in the technology. Transparency into a systems internal logic can improve trust in automation, but is hard to achieve in practice. This chapter discusses the clinical and technology related factors that influence clinician trust in automated systems, and can affect the need for transparency when developing machine learning-based clinical decision support systems.

14.1 Introduction

The recent realisation of Machine Learning (ML) techniques such as Artificial Neural Networks (ANNs) as a viable technology outside academia has opened new areas of human activity to automation. In healthcare, where human error is a significant cause of morbidity and mortality, these new approaches have revived interest in building intelligent Clinical Decision Support Systems (CDSS). Intelligent CDSS is intended as an advanced cognitive or perceptual tool to support clinicians in making sense of large amounts of data, or detecting abnormalities in complex images.

The potential applications for Machine Learning-based Clinical Decision Support (ML-CDSS) are manifold: studies have shown ANNs perform above clinicians at tasks involving interpretation of clinical data, such as diagnosing pulmonary emboli or predicting which patients are at high risk for oral cancer [23, 40]. ANNs are particularly effective in image recognition and have been applied to several radiological imaging methodologies, such as early detection of breast cancer in mammograms,

C. Gretton (✉)
University College London, School of Management, Level 38, 1 Canada Square, London E14 5AA, UK
e-mail: cosima.gretton@ucl.ac.uk

J. Zhou and F. Chen (eds.), *Human and Machine Learning*, Human–Computer Interaction Series, https://doi.org/10.1007/978-3-319-90403-0_14

with accuracy as high as 97% [35]. Humans perform particularly poorly at this task, making it ideal for automated support: some studies estimate human error rates in radiological image interpretation to be as high as 30% [4].

But technology plays a role both in preventing and unfortunately contributing to medical error [3]. New technologies impact the user and the entire system of care by altering workflows, processes, and team interactions [10]. In fast-paced inpatient environments, where multiple teams visit and give opinions on a patient, a false diagnosis by a decision support system can take on *diagnostic momentum*. Subsequent teams are less likely to question the information and will continue an already initiated treatment course [8]. This is particularly true of technologies that fully or partly automate human tasks, inducing phenomena in human operators known as automation bias and automation complacency [17]. This is the propensity for the human operator to favour the decision made by the system over their own internal judgement, or the presence of contradictory information.

ML-CDSS may be particularly at risk for inducing automation bias, posing the threat of new, unintended errors. In part this is because these models often lack transparency into their internal logic, rendering them impervious to inspection or understanding of root cause. Second, such models often find new insights and patterns in super-human amounts of data, which may prevent clinicians from evaluating the veracity of their output because the insights are novel. This has meant that despite hubris from industry there is much hesitation amongst clinicians to adopt these systems [23].

Given the current scale of medical error this is hardly surprising. In the US estimates range from 44,000 to as high as 251,454 deaths per year [25, 33], placing medical error as the third leading cause of death in the US. There is much controversy surrounding these estimates, and the lack of clarity only serves to highlight the inadequate reporting of errors in clinical medicine [49, 52]. In a system of such complexity and risk, the introduction of new technologies must be carefully considered.

There are several clinical and technology factors that can increase the likelihood of automation bias, including lack of transparency. But transparency is hard to achieve with some ML approaches and may lead to more confusion in a non-technical user. Given the challenges in developing transparent ML, optimising other clinical and technology factors may reduce the risk of automation errors and thereby the degree of transparency needed. This chapter discusses automation bias and complacency and proposes a conceptual model for the factors influencing the appropriate use of an ML-CDSS as a basis for further research.

When discussing CDSS, this chapter focusses on point-of-care systems defined as *"computer systems designed to impact clinician decision making about individual patients at the point in time that these decisions are made"*[6]. Machine learning approaches have great potential in public health and reimbursement applications but these are not considered in this chapter since they do not drive point-of-care decision-making.

14.2 Learning from History: Trends in Clinical Decision Support

Attempts to build clinical decision support systems date back half a century and provide rich insight into contextual constraints facing new ML-CDSS. The first paper on mathematical models for medical diagnosis was published in 1959 and since then attempts to automate aspects of clinical practice have followed summers and winters of artificial intelligence research [28].

Initial approaches focused on developing 'expert' diagnostic systems. These systems provided only one suggestion that the clinician was expected to follow. They focused on providing the right information at the right time: such as drug allergy alerts, guideline suggestions or diagnostic screening reminders. The rules on which they were based were relatively simple and human-understandable. For example, a colon cancer screening reminder generated when consulting a patient over a specified age [43]. The systems comprised of a knowledge base with IF/THEN rules, an inference engine with which to combine the knowledge base with patient specific data and a communication mechanism to relay the output to the clinician [50].

But in the early 1980s developers realised physicians were not interested in using these Greek oracle-like expert systems: they valued their own expertise and autonomy as decision makers. From this emerged decision support, a more collaborative approach in which a list of options is presented to the clinician. This remains the dominant approach today [38].

The decision support systems of the last century relied on an internal knowledge base. These have since evolved into non-knowledge based systems that employ machine learning or other statistical pattern recognition techniques [6, 36]. These new approaches have several advantages. Decisions in clinical practice are often made based on incomplete information. Previous knowledge-based systems perform poorly with incomplete data, but based on their training machine learning algorithms can infer missing data points and perform under uncertainty [34]. Additionally, rather than having a knowledge base derived from medical literature in need of constant updating, such systems derive associations from patient data to generate a diagnosis [6]. While this is clearly an advantage these approaches can be subject to their own unique performance limitations, which can present interpretation challenges for the clinician.

14.3 Over-Reliance and Under-Reliance in Automated Systems

As all humans, clinicians are not often aware of their own propensity for thinking errors known as cognitive biases, and have been shown to suffer from over-confidence in their abilities [5].

Table 14.1 Interaction between system performance and user response

User response	System performance			
	True positive	False positive	True negative	False negative
Agree	Appropriate reliance	Commission errors	Appropriate reliance	Omission errors
Disagree	Under-reliance	Appropriate reliance	Under-reliance	Appropriate reliance

As much as clinicians fail to recognise their own internal thinking errors, they also fail to detect the influence that technology or system design can have upon their behaviour. Technology can change behavior and induce error by occupying valuable cognitive resources through poor user interface design, poor adaptation to the clinician's workflow, or inducing automation bias [12].

Of specific relevance to ML-CDSS is automation over-reliance, a phenomenon that occurs when a human places inappropriate trust in an automated system. This takes two forms: a commission error known as *automation bias*, where the human acts upon a system's incorrect diagnosis, and an omission error known as *automation complacency*, where the system fails to make a diagnosis, and the clinician fails to spot the miss [12]. Automation bias and complacency result from the interaction of system performance and user response (see Table 14.1). There are several examples from traditional CDSS in the literature. Lyell and colleagues found that even simple e-prescribing decision-support led to automation bias. Although a correct suggestion by the CDSS reduced omission errors by 38.3%, when incorrect it increased omission errors by 33.3% [32]. Similar results were found in a study of Computer-Aided Detection (CAD) of breast cancers in mammograms: human sensitivity was significantly lower in the CAD supported condition due to errors of omission [1].

Automation under-reliance, where the human fails to trust a reliable system is also a source of error. What is clear is that for optimal human-machine performance, the human must know when to trust and when not to trust the system. Transparency influences the appropriate attribution of trust by providing insight into how the system arrived at its decision. An expert human can then evaluate the decision against their own internal knowledge [20, 46]. Evidence from other industries shows trust in recommender systems and decision support systems is increased when the system provides an explanation for its recommendation [13]. But transparency is only one factor to influence appropriate attribution of trust, and the degree to which it is needed varies depending on the context. The successful adoption of ML-CDSS in clinical practice will depend upon designing the system to elicit appropriate trust, either through transparency or other means.

14.4 Clinical and Technology Factors in Human-Machine Performance

Transparency influences trust and appropriate system use by providing insight into how the machine arrived at a decision. But full transparency is unlikely to be useful or understandable and may worsen human-machine performance. Clinicians may not be familiar with the statistical techniques underlying the technology and must use these systems under time pressure and high cognitive load. Given the heterogeneity of clinical practice transparency may also mean different things in different contexts and should be tailored to the specific goals of the human at that time. This section describes clinical and technology factors important in designing ML-CDSS for appropriate trust, and proposes a conceptual model as the basis for further research. These factors will influence trust in the system, the degree to which transparency will be important, and shape the ultimate product requirements for optimal human-machine performance.

14.4.1 Clinical Considerations for ML-CDSS

When designing point-of-care ML-CDSS there are two important clinical factors to consider that will affect the degree of transparency needed: clinical risk and the availability of expert evaluation.

14.4.1.1 Clinical Risk

The clinical risk presented by an ML-CDSS decision may influence the level of transparency needed. The United Kingdom's National Patient Safety Agency defines clinical risk as *"the chance of an adverse outcome resulting from clinical investigation, treatment or patient care."* Clinical risk can be understood in terms of severity of a healthcare hazard multiplied by the probability that it will occur [41]. For example, consider an ML-CDSS that takes real-time physiological data from a patient under anaesthesia to support the anaesthetist in titrating sedation. The impact of an error is clearly significant (high severity). Given the time pressure and operator cognitive load the probability that an error will go unidentified by the clinician and ultimately impact the patient is potentially high (high probability). In this context, system transparency around performance and the inputs on which it is basing its decision are important to enable the clinician to evaluate its output and mitigate the risk.

Contrast this with an algorithm that uses health record data to predict which members of a primary care physician's patient cohort might develop diabetes in the next five years. The clinical risk presented by an error in this example is lower: immediate interventions based on the information are minor and errors would have a

low impact (low severity). The clinician also has ample time to evaluate the validity of the decision, check orthogonal data or discuss with her colleagues. The probability that errors will go unidentified and impact the patient is lower, the clinical risk is lower and transparency may be less of a critical requirement.

Assessing the impact and probability of an error is important in defining the requirements for ML-CDSS systems. Doing so requires close collaboration with the clinicians who will ultimately be using the technology.

14.4.1.2 Expert Evaluation

The clinician plays an important role in verifying the output of an ML-CDSS and in doing so, mitigating the risk. There are several factors, including transparency that influence a clinician's ability to evaluate the output of an ML-CDSS: experience with CDSS, time pressure, interruptions, the availability of orthogonal data, familiarity with the subject matter and task complexity [17, 20, 29, 53]. Given the constraints on achieving transparency with some ML approaches such as ANNs, designers and developers may be able to optimise for other factors to elicit appropriate trust in their systems.

Experience with CDSS

In a meta-analysis of effect modifiers of automation bias, experience with a CDSS was found to decrease automation bias [20]. Repeated use of a CDSS allows a user to understand the limits of its performance and know when to place appropriate trust. Trust is one of the most extensively studied and strongest factors to affect automation bias [17]. But one of the challenges inherent in healthcare as opposed to other industries such as aviation, is the lack of reliable feedback loops. When an error occurs it might have no immediate consequences, significant time can elapse before it is discovered, or news of the error may never get back to the decision-maker [14]. In human diagnostic performance, this results in a cognitive bias called the *feedback sanction* and subsequent over-confidence in diagnostic performance [14]. In the context of human-machine interaction lack of feedback makes it hard for the user to assess the performance of a system, and thereby calibrate their trust. Feedback may vary depending on the context: in the examples above, the anaesthesiologist has immediate feedback from the physiology of the patient. The primary care physician, however, may not know if the system is correct for several years, meaning experience may not improve human-machine performance.

Experience and training also help users generate correct conceptual models of the way the system works. A conceptual model is a mental model of how a system or device works. In the absence of correct conceptual models humans form their own often erroneous conceptual models, leading to errors in using the device [42]. Even a highly simplified conceptual model can improve trust and appropriate use of a technology.

Subject matter expertise and task complexity

Familiarity with the subject matter also affects a clinician's ability to evaluate the output of a system: those less confident in their own abilities are more likely to be subject to automation bias [15]. This is closely related to task complexity and work load, which was found to be associated with automation over-reliance [21]. More experienced clinicians are likely to cope well with more complex work-loads, and potentially be better at evaluating the output of a CDSS.

The attraction of ML-CDSS lies in the potential to take large data sets and identify novel associations or predictions. For example, a 2015 paper by researchers at the Icahn School of Medicine at Mount Sinai applied a clustering algorithm to medical record and genotype data from 11,210 individuals. They identified three sub-types of type 2 diabetes, each susceptible to different complications of the disease [31]. But before use in clinical practice these novel associations will need to be validated, and will continue to be unfamiliar to clinicians. This is a critical consideration when building ML-CDSS: is the system automating current medical practice or discovering new associations? The former will be easier to implement and transparency not as essential; the clinician can compare the output with their own internal knowledge. The latter, in addition to rigorous clinical validation, may require greater transparency to elicit trust and gain adoption. As more domains are supported by CDSS there is a risk of de-skilling, and as a result a reduction in the ability of clinicians to evaluate the performance of their systems [7, 18, 19]. As an example, some electrocardiogram (ECG) machines currently provide a suggested diagnosis, written at the top of the printed page. But doctors are encouraged to ignore the decision-support and come up with their own conclusions to ensure the skill of ECG interpretation is maintained.

Time pressure

Urgency and frequent interruptions are major barriers to proper evaluation of a decision [12]. They are also universal characteristics of inpatient working conditions: a review of the literature found that nurses can be interrupted from a task over 13 times an hour [39]. Transparent ML-CDSS in such environments must be highly context specific, provide simple, task relevant information to reduce cognitive load and make it easy to return to the task after a distraction.

High urgency also removes the opportunity to consult with colleagues or assess orthogonal data. Insufficient sampling of information has been shown to be associated with increased rates of commission errors [2]. Transparency matters too: in high pressure situations the degree to which the CDSS can provide an explanation for its decisions will impact the appropriate attribution of trust in the system [37]. The primary care physician described above has ample opportunity to discuss the output of the algorithm with colleagues and decide whether to act upon its recommendations. The anaesthesiologist does not have that opportunity: the system must be sufficiently transparent for her to decide whether to trust its output without additional data or team support.

Individual differences

While not specific to clinical practice, individual differences in cognition and personality can also affect a clinician's propensity for automation bias and therefore the level of transparency that might be required [20]. Some users have a predisposition to trust an automated system, while others are more likely to distrust it [17]. In designing systems to scale across multiple clinical contexts it is hard to account for individual differences, but it is important to consider this when interpreting user feedback. Each physician may respond differently to an ML-CDSS, making it helpful to work with several different users.

14.4.2 Technical Considerations for ML-CDSS

In addition to clinical and contextual factors the design and performance of the technology itself influences trust and the likelihood of error. There is an interaction between user interface design and system performance: poor system performance and poor user interface design create a perfect storm for the inappropriate attribution of trust. The system performs poorly and the user is unable to identify the error [30]. But even a system with excellent performance can facilitate errors or bias physician behaviour if the user interface design is inadequate.

14.4.2.1 User Interface Design

Human-machine interaction errors due to user interface design are likely to be more common in healthcare than is currently known. A study of errors over a four-year period in a tertiary care hospital in Hong Kong found that 17.1% of all incidents reported were technology related, and of those 98.1% were socio-technological. The errors were not due to a technology failure, but due to how the system was operated by the user [48].

User interface design is essential for communicating system performance. Consider the following example from a device designed to deliver radiation therapy, the Therac-25, used between 1985 and 1987. It was discovered that the user interface made it possible for a technician to enter erroneous data, but despite appearing to correct it on the display the system would continue to deliver the wrong level of radiation. The only indication the dose being delivered was incorrect was an ambiguous 'Malfunction 54' code [30]. During the two years that the fault remained undiscovered multiple patients received lethal levels of radiation. Further software problems were found, each alerting the clinician by similar ambiguous malfunction codes. One technician reported 40 codes in a day, none of which enabled the clinician to understand the underlying issue [30]. Clear communication in the user interface as to the nature of the error would have avoided the continued use of this device, and continued patient harm.

User interface and information design can not only be a cause of error, but also greatly influence clinician decision-making behaviour, for better or worse. Torsvik and colleagues found that different data visualisations influenced medical students' interpretation of identical clinical chemistry results, to the extent that for one of the visualisations the results were more likely to be interpreted as within range [51]. Another study found that user interface design can directly affect treatment decisions. Persil and colleagues showed that simple grouping of antibiotic options could influence whether the clinicians chose a conservative versus an aggressive treatment for a patient with pneumonia [44]. This is important when considering how to present the output of an ML-CDSS to a clinician. Care must be taken not to inadvertently bias clinician decision-making.

14.4.2.2 System Performance

One path to reducing errors of omission and commission and improving human-machine performance is to improve system performance. As shown in Table 14.1 automation bias and complacency both occur when a system under-performs. Developers and clinicians should be aware of errors particular to statistical pattern recognition techniques that may impact performance. *Data leakage* is a phenomenon that occurs when a variable included in the training/test set contains more information than one would have access to in practice. The model exploits the variable, resulting in good performance on the test set but poor performance in practice [24]. As an example, a 2008 ML competition for detecting cancer in mammograms involved training a model on a data set which contained amongst other data points, patient IDs. The patient IDs had been assigned consecutively in the data sets, which meant the IDs were relied upon to determine the source of the data and thereby increased predictive power. But in practice, patient IDs are random and by relying on this data in training the algorithm would perform sub-optimally in the wild [47].

A similar example is that of *dataset shift*: this refers to when the conditions under which the model is trained differ from the conditions under which it is deployed. An image recognition model trained on a set of images under controlled light conditions, might fail when deployed in practice on images under varying light conditions [45]. To mitigate automation bias and complacency, systems should state performance characteristics, population demographics on which the algorithm was trained, and the conditions under which the system performs poorly [6].

14.5 The Interaction of Clinical and Technology Factors in the Attribution of Appropriate Trust

Figure 14.1 outlines a proposed conceptual model for understanding the factors that influence appropriate system use. This model is by no means exhaustive and serves to structure further discussion and investigation. Both an understanding of system

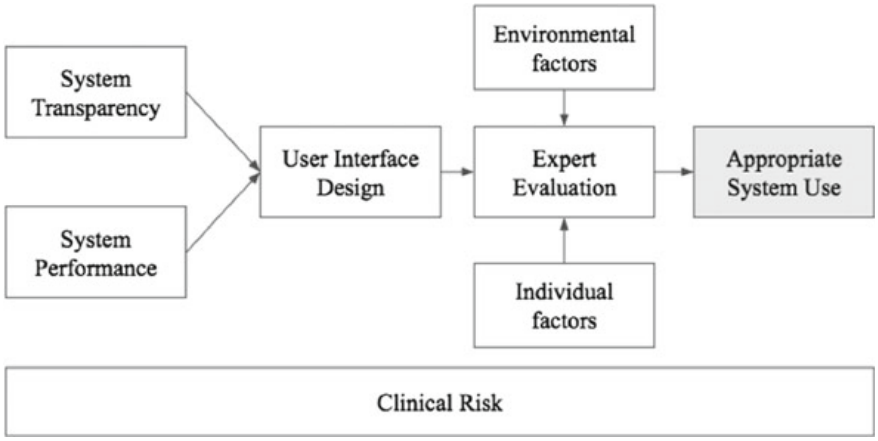

Fig. 14.1 Clinical and technology factors impacting appropriate system use

performance and transparency into the system's internal logic will help users place appropriate trust in an ML-CDSS. Both, however, depend upon good user interface design to communicate effectively to the user.

Clear communication of this information will enable expert evaluation. Expert evaluation is itself determined by individual (e.g. training) and environmental (e.g. time pressure) factors. Clinical risk is important throughout, influencing every consideration from acceptable performance characteristics to the need for regular reviews of appropriate utilisation once the system is in routine use.

14.6 Adoption of ML-CDSS: Legal, Ethical and Policy Implications Beyond Point-of-Care

Designing point-of-care systems with attention to the factors described may improve system design and reduce the risk of error. But transparency into an ML-CDSS's internal logic is important beyond the bedside.

One of the major concerns regarding the lack of transparency, which cannot be addressed through other means, lies in the attribution of blame in situations of medical error. Technology developers often place the burden of responsibility on the clinician [26]. The clinician must use the device within the bounds of their medical knowledge and interpret the information in the context of the patient. The Therac-75 case highlights how poor communication and lack of transparency limits the information available, meaning the user cannot make an informed decision [16, 30]. To justifiably defer responsibility, the technology must equip the clinician with sufficient information to make an informed decision. Further, transparency is essential for identifying root cause and attributing blame. This concern is reflected in a recent directive from the European Union states that by 2018, companies deploying algorithms that influence the public must provide explanations for their models' internal logic [22].

Second, from an ethical and legal standpoint transparency is needed to support clinicians in gaining informed consent. If the physician does not understand the logic behind a certain treatment recommendation they cannot reasonably inform the patient and obtain consent.

Finally, true adoption in medicine depends on obtaining clinical utility data and updating medical guidelines. For algorithms that generate novel associations transparency may be needed in order for policy makers and medical societies to trust the findings and invest in costly clinical trials or health economic studies.

14.7 Conclusion

Clinical decision support in medicine has a rich history and is undergoing a renaissance with the advent of new machine learning techniques. But new technologies face the same challenges as the previous approaches. The inappropriate attribution of trust is one of the major barriers to widespread adoption and leads to medical error in the form of omission and commission errors. Lack of transparency is an issue for clinical practice as it prevents physicians from evaluating decision-support outputs against their own internal knowledge base. But full transparency, given the conditions under which clinicians work, is hard to achieve and may negatively impact trust. Different degrees of transparency may be needed depending on clinical risk and the ability of the expert to evaluate the decision. Designing with an appreciation of real-world practice constraints such as time pressure, combined with good user interface design to enable expert evaluation can facilitate appropriate use. Early engagement with clinicians in the design, development and implementation of new technologies will reduce risks and improve system adoption [27]. Given the potential for new errors and work-arounds, continued monitoring of technologies as they enter common use is important to ensure patient safety [9].

These practical, ethical and legal constraints on ML-CDSS may mean that developers are forced to take different approaches if ML-techniques are unable to provide the required transparency [11]. But medicine is highly heterogeneous and local collaborations between clinicians and technologists will identify niche areas where risk, transparency and utility align and ML-based approaches can provide value.

References

1. Alberdi, E., Povyakalo, A., Strigini, L., Ayton, P.: Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography. Academic Radiology **11**(8), 909–918 (2004)
2. Bahner, J.E., Hüper, A.D., Manzey, D.: Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. International Journal of Human Computer Studies **66**(9), 688–699 (2008)
3. Battles, J.B., Keyes, M.A.: Technology and patient safety: A two-edged sword (2002)

4. Berlin, L.: Radiologic errors, past, present and future. Diagnosis **1**(1), 79–84 (2014)
5. Berner, E.S., Graber, M.L.: Overconfidence as a Cause of Diagnostic Error in Medicine. American Journal of Medicine (2008)
6. Berner, E.S., La Lande, T.J.: Overview of Clinical Decision Support Systems. In: Clinical Decision Support Systems, pp. 1–17. Springer, Cham (2016)
7. Berner, E.S., Maisiak, R.S., Heudebert, G., Young, K.: Clinician performance and prominence of diagnoses displayed by a clinical diagnostic decision support system (2003)
8. Campbell, S.G., Croskerry, P., Bond, W.F.: Profiles in Patient Safety: A "Perfect Storm" in the Emergency Department. Academic Emergency Medicine **14**(8), 743–749 (2007)
9. Carayon, P., Kianfar, S., Li, Y., Xie, A., Alyousef, B., Wooldridge, A.: A systematic review of mixed methods research on human factors and ergonomics in health care (2015)
10. Carayon, P., Schoofs Hundt, A., Karsh, B.T., Gurses, A.P., Alvarado, C.J., Smith, M., Flatley Brennan, P.: Work system design for patient safety: the SEIPS model. Quality and Safety in Health Care **15**(suppl–1), i50–i58 (2006)
11. Castelvecchi, D.: Can we open the black box of AI? Nature **538**(7623), 20–23 (2016)
12. Coiera, E.: Technology, cognition and error. BMJ Quality & Safety **24**(7), 417–422 (2015)
13. Cramer, H., Evers, V., Ramlal, S., Van Someren, M., Rutledge, L., Stash, N., Aroyo, L., Wielinga, B.: The effects of transparency on trust in and acceptance of a content-based art recommender. User Modeling and User-Adapted Interaction **18**(5), 455–496 (2008)
14. Croskerry, P.: The feedback sanction. Academic Emergency Medicine **7**(11), 1232–8 (2000)
15. Dreiseitl, S., Binder, M.: Do physicians value decision support? A look at the effect of decision support systems on physician opinion. Artificial Intelligence in Medicine **33**(1), 25–30 (2005)
16. Dworkin: Autonomy and informed consent. President's Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioral Research Making Health Care Decisions. Federal Register **3**(226), 52,880–52,930 (1982)
17. Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., Beck, H.P.: The role of trust in automation reliance. International Journal of Human Computer Studies **58**(6), 697–718 (2003)
18. Friedman, C.P., Elstein, A.S., Wolf, F.M., Murphy, G.C., Franz, T.M., Heckerling, P.S., Fine, P.L., Miller, T.M., Abraham, V.: Enhancement of Clinicians' Diagnostic Reasoning by Computer-Based Consultation: A Multisite Study of 2 Systems. JAMA **282**(19), 1851–1856 (1999)
19. Friedman, C.P., Gatti, G.G., Franz, T.M., Murphy, G.C., Wolf, F.M., Heckerling, P.S., Fine, P.L., Miller, T.M., Elstein, A.S.: Do physicians know when their diagnoses are correct? Implications for decision support and error reduction. Journal of General Internal Medicine **20**(4), 334–339 (2005)
20. Goddard, K., Roudsari, A., Wyatt, J.C.: Automation bias: a systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association **19**(1), 121–127 (2012)
21. Goddard, K., Roudsari, A., Wyatt, J.C.: Automation bias: Empirical results assessing influencing factors. International Journal of Medical Informatics (2014)
22. Goodman, B., Flaxman, S.: EU regulations on algorithmic decision-making and a "right to explanation". 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016) (Whi), 26–30 (2016)
23. Holst, H., Aström, K., Järund, A., Palmer, J., Heyden, A., Kahl, F., Tägil, K., Evander, E., Sparr, G., Edenbrandt, L.: Automated interpretation of ventilation-perfusion lung scintigrams for the diagnosis of pulmonary embolism using artificial neural networks. European journal of nuclear medicine **27**(4), 400–406 (2000)
24. Kaufman, S., Rosset, S.: Leakage in data mining: Formulation, detection, and avoidance. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining pp. 556–563 (2012)
25. Kohn, Linda T.; Corrigan, Janet M.; Donaldson, M.S.: [To err is human: building a safer health system]., vol. 21 (2002)
26. Koppel, R., Kreda, D.: Health Care Information Technology Vendors' Hold Harmless Clause. JAMA **301**(12), 1276–1278 (2009)

27. Korunka, C., Weiss, A., Karetta, B.: Effects of new technologies with special regard for the implementation process per se. Journal of Organizational Behavior **14**(4), 331–348 (1993)
28. Ledley, R.S., Lusted, L.B.: Reasoning foundations of medical diagnosis; symbolic logic, probability, and value theory aid our understanding of how physicians reason. Science (New York, N.Y.) **130**(3366), 9–21 (1959)
29. Lee, C.S., Nagy, P.G., Weaver, S.J., Newman-Toker, D.E.: Cognitive and system factors contributing to diagnostic errors in radiology (2013)
30. Leveson, N.G., Turner, C.S.: An Investigation of the Therac-25 Accidents. Computer **26**(7), 18–41 (1993)
31. Li, L., Cheng, W.Y., Glicksberg, B.S., Gottesman, O., Tamler, R., Chen, R., Bottinger, E.P., Dudley, J.T.: Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Science Translational Medicine **7**(311), 311ra174–311ra174 (2015)
32. Lyell, D., Magrabi, F., Raban, M.Z., Pont, L., Baysari, M.T., Day, R.O., Coiera, E.: Automation bias in electronic prescribing. BMC Medical Informatics and Decision Making **17**(1), 28 (2017)
33. Makary, M.A., Daniel, M.: Medical error the third leading cause of death in the US. BMJ p. i2139 (2016)
34. Marakas, G.: Decision Support Systems in The 21st Century. Prentice Hall (1999)
35. Mehdy, M.M., Ng, P.Y., Shair, E.F., Saleh, N.I.M., Gomes, C.: Artificial Neural Networks in Image Processing for Early Detection of Breast Cancer. Computational and Mathematical Methods in Medicine **2017**, 1–15 (2017)
36. Metzger, J., MacDonald, K.: Clinical decision support for the independent physician practice. October (2002)
37. Miller, R.A., Gardner, R.M.: Summary recommendations for responsible monitoring and regulation of clinical software systems. Annals of Internal Medicine **127**(9), 842–845 (1997)
38. Miller, R.A., Masarie, F.E.: The demise of the 'Greek Oracle' model for medical diagnostic systems. Methods of Information in Medicine **29**(1), 1–2 (1990)
39. Monteiro, C., Avelar, A.F.M.: Pedreira, M.d.L.G.: Interruptions of nurses' activities and patient safety: an integrative literature review. Revista Latino-Americana de Enfermagem **23**(1), 169–179 (2015)
40. Naguib, R.N.G., Sherbet, G.V.: Artificial neural networks in cancer diagnosis, prognosis, and patient management (2001)
41. National Patient Safety Agency: Healthcare risk assessment made easy. National Patient Safety Agency **3**(March) (2007)
42. Norman, D.a.: The Design of Everyday Things: Revised and Expanded Edition (1988)
43. Osheroff, J.A.: Improving Medication Use and Outcomes with Clinical Decision Support:: A Step by Step Guide. HIMSS (2009)
44. Persell, S., Friedberg, M., Meeker, D., Linder, J., Fox, C., Goldstein, N., Shah, P., Doctor, J., Knight, T.: Use of behavioral economics and social psychology to improve treatment of acute respiratory infections (BEARI): rationale and design of a cluster randomized controlled trial [1RC4AG039115-01] - study protocol and baseline practice and provider characteris. BMC infectious diseases **13**, 290 (2013)
45. Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.: Dataset Shift in Machine Learning. MIT Press (2008)
46. Rogers, Y., Rogers, Y: A brief introduction to Distributed Cognition. Cognitive Science (1997)
47. Rosset, S., Perlich, C., Świrszcz, G., Melville, P., Liu, Y.: Medical data mining: Insights from winning two competitions. Data Mining and Knowledge Discovery **20**(3), 439–468 (2010)
48. Samaranayake, N.R., Cheung, S.T., Chui, W.C., Cheung, B.M.: Technology-related medication errors in a tertiary hospital: A 5-year analysis of reported medication incidents. International Journal of Medical Informatics **81**(12), 828–833 (2012)
49. Shojania, K.G., Dixon-Woods, M.: Estimating deaths due to medical error: the ongoing controversy and why it matters: Table1. BMJ Quality & Safety pp. bmjqs–2016–006,144 (2016)
50. Tan, J., Sheps, S.: Health decision support systems (1998)
51. Torsvik, T., Lillebo, B., Mikkelsen, G.: Presentation of clinical laboratory results: an experimental comparison of four visualization techniques. Journal of the American Medical Informatics Association **20**(2), 325–331 (2013)

52. Weingart, S.N.: McL Wilson R, R.M., Gibberd, R.W., Harrison, B.: Epidemiology of medical error. The. Western journal of medicine **172**(6), 390–3 (2000)
53. Westbrook, J.I.: Association of Interruptions With an Increased Risk and Severity of Medication Administration Errors. Archives of Internal Medicine **170**(8), 683 (2010)