Computers & Graphics 128 (2025) 104212

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Special Section on Highlights from EuroVA 2024

Trust at every step: Embedding trust quality gates into the visual data exploration loop for machine learning-based clinical decision support systems

Dario Antweiler *, Georg Fuchs

Fraunhofer IAIS, Sankt Augustin, Germany

ARTICLE INFO

Keywords:
Healthcare
Visual analytics
Trustworthiness
Machine learning

ABSTRACT

Recent advancements in machine learning (ML) support novel applications in healthcare, most significantly clinical decision support systems (CDSS). The lack of trust hinders acceptance and is one of the main reasons for the limited number of successful implementations in clinical practice. Visual analytics enables the development of trustworthy ML models by providing versatile interactions and visualizations for both data scientists and healthcare professionals (HCPs). However, specific support for HCPs to build trust towards ML models through visual analytics remains underexplored. We propose an extended visual data exploration methodology to enhance trust in ML-based healthcare applications. Based on a literature review on trustworthiness of CDSS, we analyze emerging themes and their implications. By introducing trust quality gates mapped onto the Visual Data Exploration Loop, we provide structured checkpoints for multidisciplinary teams to assess and build trust. We demonstrate the applicability of this methodology in three real-world use cases – policy development, plausibility testing, and model optimization – highlighting its potential to advance trustworthy ML in the healthcare domain.

1. Introduction

Machine learning (ML) presents unique opportunities to lower healthcare costs [1], reduce disease burden [2], and enhance patient outcomes [3] as well as improve satisfaction for both patients and healthcare staff [4]. Consequently, its ability to analyze large datasets plays a crucial role in the ongoing digital transformation of healthcare systems. Among the many applications of ML in healthcare, Clinical Decision Support Systems (CDSS) stand out due to their vast potential and exceptionally high trust requirements. Example applications include risk prediction [5], radiology diagnostics [6] and early detection of pulmonary hypertension [7]. Lack of trust is one of the main reasons why, despite extensive efforts, only a limited number of such projects have been effectively implemented in real-world clinical settings [8]. Visual analytics (VA) is widely regarded as a means to enhance trust in automated decision systems, serving as an essential tool for both end-users and the model development phase of these systems. In the case of ML, this comprises versatile interactions between users and models via adaptable visualizations. Visual analytics can assist both healthcare professionals (HCPs) and data scientists during joint model development.

Our contribution is a structured review of the topic and the development of a general methodology, rather than a specific instantiation. We propose an extended visual data exploration methodology towards trustworthy ML in healthcare. Building upon our previous work [9], our added contributions are as follows: First, we report on an extensive literature review of 62 publications on the topic of trustworthy CDSS. Second, we structure and analyze emerging themes and carve out implications. Third, we map the identified trustworthiness aspects onto the visual data exploration loop, a generic process of data exploration and modeling supported by interactive visualizations [10]. Fourth, we introduce the concept of trust quality gates, which define quality criteria to be established before development and verified during implementation. Finally, we exemplify the applicability of our approach by discussing three concrete case studies across different healthcare domains.

2. Common pitfalls

Our goal for the methodology is to avoid common pitfalls that are prevalent during model development for CDSS regarding trust. We identified these pitfalls during multiple research and industry projects in diverse healthcare settings:

* Corresponding author.
E-mail addresses: dario.antweiler@iais.fraunhofer.de (D. Antweiler), georg.fuchs@iais.fraunhofer.de (G. Fuchs).

https://doi.org/10.1016/j.cag.2025.104212
Received 15 January 2025; Received in revised form 11 March 2025; Accepted 12 March 2025
Available online 2 April 2025

Pitfall 1: Failing to consider trust and trustworthiness or assuming them to be self-evident. Often during development, it is assumed that trust automatically arises with usage over time or is simply not considered at all. Contrary to that, research shows that trust is paramount to successful implementations of CDSS.

Pitfall 2: Confusing trust with trustworthiness. Trust is a subjective feeling, whereas trustworthiness represents measurable properties of a system. It should not be viewed purely from a psychological perspective. Being trustworthy contributes to earning trust, but it does not guarantee trust on its own [11].

Pitfall 3: Viewing trust and trustworthiness as unimodal. Trust is not achieved solely through the quality of output results. Instead, trust is much more complex and has many aspects including fairness, ease of use and autonomy.

Pitfall 4: Considering trust and trustworthiness isolated across components and stakeholders. Trust flows throughout the entire process and must be considered from the outset, not just at the product finalization stage. Often, clinicians and data scientists think separately and use different terminologies, which undermines the development of trust across the process.

Pitfall 5: Not employing visual analytics for trust and trustworthiness. Visual analytics has untapped potential to enhance trust and trustworthiness, supporting both trust building and calibration.

To circumvent these pitfalls, we first identify aspects of trust through analysis of related work and aggregate reasons for failures and successes of ML-based CDSS implementations. We subsequently cluster these aspects and map them onto the visual data exploration loop.

3. Related work

3.1. Trust and trust models

The concept of trust has been studied in various research fields, including sociology, psychology and economics [12]. It is described as a relationship between a *trustor* and a *trustee*, based on properties that make the trustee *trustworthy* [13]. *Trustworthiness* is distinguished from trust by being objectively verifiable, while trust is formed as a belief within the user. Trust-building encompasses the dynamic processes and interactions that actively foster trust among users over time. Our focus lies on decision-making processes that rely on trust. Previous works structures the analysis of contextualized trust problems during decisions [14]. We build on their concept of separating *claims*, *confidences* and *evidences* during our analysis. [15]

3.2. Trustworthy ML through visual analytics

Although highly important, the conceptualization of machine learning trustworthiness is still ambiguous in research and practice [16]. Trustworthiness is interconnected along all stages of the machine learning pipeline [17,18]. It can be enhanced by ensuring transparency, robustness, and fairness [19]. Practical strategies to achieve this with visual analytics methods include uncertainty visualization [20], providing clear data provenance [21], and allowing users to interact with surrogate models [22,23]. Even in novel visual analytics systems, domain experts can have high trust if the systems are designed to be intuitive, transparent, and flexible, enabling users to easily transition between various analysis tasks [24]. Interactive visualizations help build trust in machine learning systems [25]. A comprehensive overview of current methodologies is provided by recent reviews [17, 26,27].

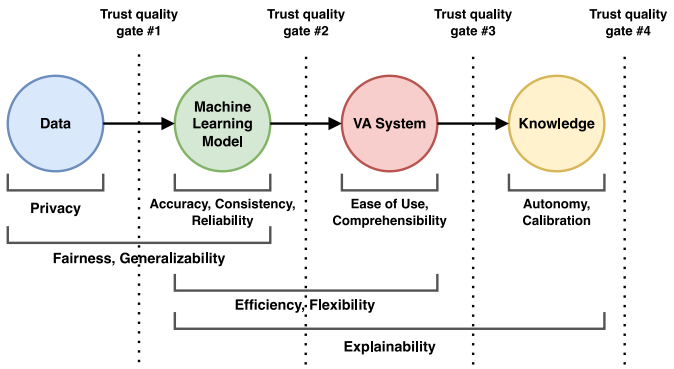

Fig. 1. Trustworthiness aspects structured into clusters along the visual data exploration loop as identified by our literature review. Dotted lines highlight trust quality gates that enable checking for sufficient trust building.

3.3. ML-based clinical decision support systems

The broad definition of CDSS encompasses various ML-based applications, including the prediction of delirium [28], sepsis [29], disease progression [30], surgery risk [31], detecting epileptic seizures [32], or tumor board evaluations [33]. They can be further categorized along the dimensions of time of use, type of trigger, information output and model architecture for inference. Analytical techniques are commonly grouped into descriptive, diagnostic, predictive and prescriptive analytics. Recent reviews provide an overview over the CDSS-landscape and highlight accompanying challenges, specifically clinical validation and adoption [34,35]. Challenges include fragmented workflows, alert fatigue as well as lack of digital literacy, data quality and interoperability [36].

Notably missing from the literature is a comprehensive methodology that specifically addresses how visual analytics can be used to systematically enhance trust in ML models for CDSS, particularly through structured checkpoints and trust quality gates integrated within the visual data exploration loop.

4. Methodology

Our original framework introduced stakeholders and their relationships as well as an analysis of the flow of trust between them [9]. Key insights included the role of visual analytics experts as trust intermediaries between data scientists and healthcare professionals, as well as the need to reduce complexity, prioritize salient factors, and incorporate domain-specific terminology relevant to the task at hand. Here, we enrich and extend this principle by introducing trust quality gates to establish structured checkpoints for multidisciplinary teams to evaluate progress and foster trust during development. This section is organized into trustworthiness aspects, interrelation between them, trust-building measures and finally trust quality gates.

4.1. Trustworthiness aspects

We structure trustworthiness aspects into areas to better understand their commonalities and differences (cf. Fig. 1). This structure is grounded in a literature review of 62 relevant publications published over the last 17 years, as seen in Fig. 2. To conduct the literature review, we searched for the keywords '*trust*' and '*CDSS*' or '*clinical decision support system*' on Google Scholar. Additionally, we employed backward citation tracking by reviewing references in the identified works. Only publications that explicitly addressed trust within the article were selected for inclusion. Additionally, many pairs of trustworthiness aspects are interrelated and should be examined together. We will highlight such interrelations and their consequences.

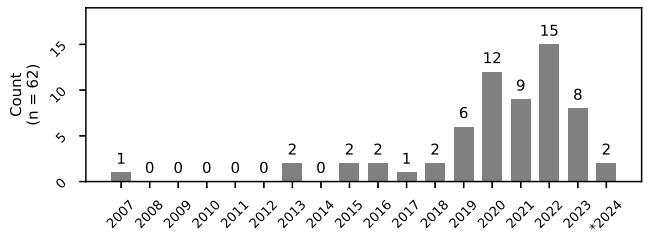

Fig. 2. Statistic of trust-related CDSS-publications identified and analyzed in the literature review over time. An uptake in recent years is clearly visible.

① *Data Quality & Privacy.* One of the foundational aspects of trustworthiness of any data-driven system is data quality, which includes the accuracy, consistency, completeness, and reliability of data. Perceived low data quality can create mistrust in the overall CDSS [37,38], and for good reason. It is well established that data quality has a larger impact on the resulting model than model tweaking or even model choice [39]. Low-quality data is very likely to negatively impact a model's accuracy and reliability, as well as increasing the risk of insufficient fairness and generalizability.

Health data is almost always related to individuals and thereby protected by dedicated privacy laws such as HIPPA or GDPR, by which organizations are bound to comply with legal regulations surrounding data collection, storage, and use [40,41]. Privacy compliance goes beyond purely legal implications, however. Anonymization methods such as *k-anonymity* or *ϵ-differential privacy* may artificially degrade data sets, subsequently reducing the model's quality [42]. On the other hand, respecting professional discretion by protecting against data leakage is paramount, as it was repeatedly shown to be possible to extract data from trained models [43], and privacy breaches are a key concern of clinicians regarding their trust in CDSS [44]. Thus, privacy measures and their potential impacts need to be communicated to CDSS users using appropriate visual indicators [45].

② *Accuracy, Reliability & Consistency.* Complement to data quality considerations, quantitative validation of employed algorithms is paramount for trust building in healthcare settings [46]. High accuracy as well as high model confidence improves trustworthiness of a system [47,48]. Inaccurate advice in general and excessive false positives in particular reduce trustworthiness and clinical meaningfulness [49,50]. Too low recall or a high false negative rate can even be an objective health risk. Those aspects are due to two types of uncertainty: Aleatoric uncertainty arises from inherent randomness or noise in the data and cannot be reduced by gathering more data. Epistemic uncertainty on the other hand refers to incomplete modeling or data [51]. Examining regional and demographic variability of training data can guarantee fair, robust validation, enhancing the reliability and suitability of ML products for clinical use [5]. Explicit documentation of the purpose of data collection and underlying assumptions, (either in advance or as a result of examination) enhances safety. In fact, acknowledging the imperfections and limits of algorithms increases the trustworthiness of machine learning systems in healthcare, especially in situations of discrepancy (lack of concordance) between a system's output and a clinician's assessment [52–55]. On the other hand, studies suggest that providing clinicians with only a model's overall accuracy is insufficient for building trust in predictive CDSSs [48]. This indicates that accuracy, reliability & consistency should be one of several elements in building trust [8].

③ *Ease of use & Comprehensibility.* Machine learning systems should have interfaces adequate for healthcare users and their respective work processes to generate trust [56]. This includes the expectation to display the right information at the appropriate level of detail at the right time to the right people. You cannot trust what you do not understand, and you can only understand what you can examine appropriately [57,58]. As expected, perceived understandability does influence clinician's trust in CDSS [48]. Usually, clinicians try to understand the logic behind software systems at least superficially. The large number of inputs and intricate model interactions involved in latest architectures can thus hamper trust building [59]. Awareness about inner workings of machine learning can be achieved through feedback and alerts, thereby raising trustworthiness [60].

On the other hand, clinicians prefer not to make decisions about trust at every junction [61]. Research indicates that the depiction of the model's inner workings may increase comprehensibility, but not user's trust [47]. A positive correlation between physician understanding and trust has been observed [62]. This seems like a contradiction at first glance, but is precisely one major reason to employ visual analytics as part of the model development process. It is of utmost relevance how the inner workings are presented, how users can interact with them, and what kind of insights are generated. Lack of understanding of how machine learning processes inputs may quickly erode trust [53]. Comprehensibility can be at odds with model accuracy, but both are required to gain user's trust [48].

④ *Autonomy & Calibration.* Clinicians should have the *lead role* during clinical decisions and be permanently made aware whether they are in contact with a machine learning system or not [63]. This aspect of autonomy extends to the possibility of justified deviations from a recommendation or "over-riding" a decision as well [57]. Additionally, outputs of machine learning systems should not be perceived or represented as "competition" to human assessments, diminishing professional roles [57,59]. Clinicians can experience overtrust or over-reliance into machine learning systems, where trust is unwarranted and which can lead to false application [41,64,65]. Professionals may place over-trust in the insights provided by these tools, either due to limited expertise in using the technology or the complexity of the cases involved [66].

⑤ *Fairness & Generalizability.* Fair, inclusive and unbiased data is a prerequisite to design "safe, explainable, reliable and trustworthy AI systems", which includes machine learning [66]. Algorithmic bias reduces machine learning trustworthiness in healthcare [38]. Some clinicians state health equity concerns with machine learning systems as a reason for slow uptake [67]. Generalizability from one hospital to another is a major quality criterion clinicians consider when deciding for or against the usage of a system. Consistent models will likely be more generalizable when trained on large amounts of data.

⑥ *Efficiency & Flexibility.* Clinicians require decision-making based on CDSS to be faster, or at least not slower, than existing processes, while not compromising safety and security. They need the ability to cope with process-related and medical varieties, as well as diverse demographics [5]. This includes the adaptation to local customs as well as the integration into existing workflows.

⑦ *Explainability.* The notion of explainability includes both transparency and oppositely model opaqueness. Transparency is essential for trustworthiness in machine learning systems [6]. We can further separate this aspect into *data opaqueness* and *model opaqueness*. Many systems are developed using secondary data that was not originally collected with the specific application in mind, which can lead to unforeseen challenges such as bias or quality issues [8]. This relates to fairness and consistency issues as well. The lack of transparency in certain machine learning models has resulted in reduced acceptance among healthcare practitioners [68]. Black box machine learning models generate more distrust than rule-based models [40]. This is especially true for complex model architectures, such as neural networks [38,69]. For users with limited domain-specific knowledge, explanations may be less essential, whereas they are crucial for those with high expertise [70]. Transparency is necessary for trust to be

formed for effective adoption [71] and clinicians even value open models higher than diagnostic or prognostic accuracy [72]. Furthermore, studies have shown that explanations can increase clinicians' confidence in their own decisions [46]. On the other hand, transparency can lead to undue trust in models, reducing the users' ability to detect mistakes [72]. Additionally, explanations can increase cognitive effort and further burden healthcare professionals and should therefore be assessed as a trade-off [46]. Transparency is often misaligned with the intention of corporate developments to protect their business secrets and intellectual property.

Trust is not only correlated to transparency regarding the model, but the operationalization of the machine learning systems as well [73]. There is additionally the need to distinguish between trust in the model as a whole versus trust in individual predictions [49]. Clinicians report a desire to understand an algorithm's decision thresholds, similarly to how they currently adjust to the tendencies of their peers [53]. The impact of explainability can be further enhanced through the design and usage of user-centered explanations [48]. Additionally, default explanations can be replaced by on-demand comparisons to standard clinical care [61].

Regulation. In many countries, AI systems in healthcare are not sufficiently regulated [67]. Lack of regulatory oversight is a risk that might outweigh potential benefits [57]. Establishing guidelines and policies enhances trust in machine learning systems [58], a lack thereof leads to uncertainties and diminishing trust [74]. Additionally, regulation is needed to prevent incentives and financial interests regarding machine learning systems and their decisions, reinforcing trust [58].

4.2. Interrelation between aspects

The identified trustworthiness aspects relate to each other in many ways and typically do not manifest themselves discretely. We localize all aspects along the visual data exploration loop to determine the relationships and interactions between them is detail.

Privacy ↔ Fairness & Generalizability On the one hand, anonymized data aims to prevent usage of protected information as features in machine learning models and thereby (presumably) increases fair outcomes. However, the anonymization process cannot rule out the possibility that models unintentionally identify *proxy features* for protected characteristics and employ them for decisions. On the contrary, this can lead to further unfair treatments, as model developers may be lulled into a false sense of security or overlook biases, when confronted with scrutinized model results. Strict privacy protections can limit the amount of available training data, thereby reducing the generalizability of trained models and limiting trustworthiness. This can be mitigated by privacy-preserving training methods, including federated learning or synthetic data generation.

Accuracy, Reliability & Consistency ↔ Fairness & Generalizability Both of these aspects are related by their goal to achieve equivalent results on different patient cases. Fairness concerns achieving consistent results across populations, ensuring uniform outcomes irrespective of protected attributes, such as gender, race, or origin. Reliability and consistency, by contrast, focus on the system's robustness against slight, non-meaningful changes in input features, so-called *perturbations*. Models that demonstrate consistency are more likely to generalize effectively when trained on extensive datasets.

Accuracy, Reliability & Consistency ↔ Explainability *Explainability* can increase *accuracy, consistency and reliability* by gauging feature relevance, identifying hidden relationships and preventing overfitting. Detailed uncertainty measures provide means to assess consistency of model outputs [75]. Explainable models are more reliable, accurate and consistent as failure modalities are easier to detect and prevent. Likewise, insights into the inner workings of a model might pinpoint data quality issues which in turn increases the model's accuracy.

Explainability ↔ Fairness & Generalizability The aspect of explainability not only addresses data and modeling quality, but also relates to *fairness* and *generalizability* issues. Transparent models enable the identification of unequal treatment of (protected) groups and therefore raised awareness for fair outcomes. At the same time, explainability leads to insights that increase generalizability, such as feature dependencies, unintended feature influence on outcomes or discrepancies. Both data scientists and healthcare professionals can use those insights to develop models that better transfer to different institutions, cohorts and medical fields.

Explainability ↔ Ease of Use/Comprehensibility Explainability is related to *comprehensibility* due to the fact that you can only understand what you can examine appropriately. This means that employed models should be explorable by the user, preferably by easy to use and versatile visual inspection tools. It includes comparison of models based on multiple benchmarks, metrics and cohorts. Explainability tools such as intermediate concepts or feature clustering can support healthcare professionals in dealing with complex models and to put results into context more easily. As a property of a model, explainability must also be effectively communicated to the user through the user interface and training (system comprehensibility). How the user ultimately utilizes this information depends on the system's ease of use.

Explainability ↔ Autonomy & Calibration Explainability increases *autonomy* of users, by providing arguments to reason for or against the execution of a system's result as well as opportunities to justify a deviation from it. These justifications are paramount in healthcare settings, both for explanations towards the patients and statements in the case of legal questions. Explanations that are similar to human reasoning enable systems to be more calibrated to their users.

Efficiency & Flexibility ↔ Ease of Use/Comprehensibility Comprehensibility of a machine learning system is directly linked to efficiency and flexibility. A flexible system that individually tailors outputs to patient cases reduces cognitive load while increasing efficiency. This allows healthcare professionals to quickly understand and implement CDSS recommendations without feeling overwhelmed by unnecessarily complex or irrelevant options. Ease of use may be at odds with flexibility, when adaptation to many different workflows creates new complexities. This needs to be alleviated by context-dependent user interfaces and visualizations. Conversely, non-intuitive systems almost always introduce inefficiencies in clinical workflows, reducing trust on both sides.

Efficiency & Flexibility ↔ Accuracy, Reliability & Consistency Additionally, more accurate, consistent and reliable models lead to more efficient and flexible tools in the clinical practice. This is because reliable models reduce the need for extensive manual oversight and ad-hoc scrutiny, allowing healthcare professionals to focus on patient care rather than troubleshooting system errors. Consistent models are easier to adapt to differing clinical scenarios without compromising performance.

4.3. Trust-building measures

Ultimately, decision makers (the trustors) must have a certain amount of trust present, whether consciously or not, before model-generated information from the CDSS (the trustee) is used in the decision-making process. The preceding sections outlined how trustworthiness aspects driving overall trust building can be captured in a structured way along the chain of actors (in the visual data exploration loop) leading up to model-based decisions [14]. At the same time, it has been found that human actors typically do not have a consistent process to develop trust in new models, and that the set and individual weight of trustworthiness aspects depends on the individual and circumstances [14,72,76]. It is therefore hardly feasible to prescribe a rigid set of trustworthiness aspects and associated measures. It is possible and useful, however, to provide a mapping from relevant trustworthiness aspects (*claims*) and their respective, task- and application-specific instantiation (*requirements*) to possible trust building measures providing *evidence* these requirements have either

Table 1
Trust quality gates with their respective claims, requirements and evidence supported by Visual Analytics methods.

Claim *Property of the data/model/system*	Requirement *Application-specific threshold with respect to claim*	Evidence *Proof, that the requirement fulfills the claim*
Trust quality gate #1		
Data explorability by healthcare professionals	Quality assessment of training data	Visual interfaces showing data quality and distribution
Protection of sensitive patient data	Compliance with privacy-preserving methods	Visual representations of privacy mechanisms
Representativeness of dataset for examined cohort and desired use case	Statistical thresholds for acceptable bias levels (e.g., subgroup representation)	Visualization of subgroup distributions and bias detection results (e.g., demographic histograms, parity heatmaps)
Appropriate level of outliers in dataset	Thresholds for acceptable outlier ratios; methods for outlier detection & handling	Visualization of outliers within data distribution
Trust quality gate #2		
High model accuracy	Accuracy thresholds specific to the application (e.g., 90% for prediction tasks)	Visualization of accuracy metrics (e.g., confusion matrices, performance curves and distributions
Alignment with clinical goals	Specific benchmarks for clinical impact (e.g., accuracy, precision, recall)	Visualization of clinical performance metrics (e.g., ROC curves)
Equitable performance across protected groups	Fairness metrics such as demographic parity, equalized odds	Visualization of statistical analysis metrics
Robustness against systematic weaknesses	Performance thresholds under adversarial perturbations or distribution shifts	Visualizing performance variations and model failures
Consistent predictions across varying scenarios	Thresholds for acceptable variability in predictions under perturbations	Visual comparison of predictions in different conditions
Model's decisions are meaningfully explained to stakeholders	Implementation of interpretability techniques	Visualization of interpretability results (e.g., feature importance, SHAP values)
Assumptions underlying the model design and development are apparent	Documentation and evaluation of modeling assumptions (e.g. linearity)	Visual comparisons of model assumptions versus real-world data
Monitoring and mitigation of model and concept drift	Thresholds for drift detection metrics (e.g., KL divergence)	Visualizations of data distributions and model performance
Outperformance of baselines and competitiveness with status quo	Performance thresholds relative to baseline models and approaches	Visual performance comparison including clinical outcome metrics
Model generalizes well to diverse populations and settings	Similar performance across different demographics and clinical scenarios	Visual comparison of performance differences
Trust quality gate #3		
Provision of the right amount of information for stakeholders to make informed decisions	Defined criteria for granularity in outputs (e.g., detailed vs. high-level summaries)	Customizable detailedness of visualizations
Intuitive and comfortable use by healthcare professionals in clinical practice	Interfaces that simplify data interpretation and decision-making	Dashboards, workflow visualizations and familiar views tailored to use case and clinical terminology
Reduced cognitive load on healthcare professionals	Reduced time-on-task, simple interactions and decision pathways	Intuitive charts and dashboards, relevance highlighting, visual workflow
System operates efficiently with computational resources and clinician time	Model inference speed and clinician task completion time	Visually contrasting inference duration of model architectures
System adapts to new tasks, workflows & datasets with minimal reconfiguration	Integration without significant retraining or redesign	Adaptable visualizations and customizable workflows
Trust quality gate #4		
Appropriate levels of human oversight and autonomy	Defined thresholds for when human intervention is required	Visualization of adjustable thresholds, alternatives, corresponding implications
Calibration to align with clinician's decision-making	Thresholds for calibration metrics of clinical judgment	Visual comparison of decision boundaries, confidences and agreements
Support of communication with model as a trustworthy actor in decision process	Explain decisions, highlight uncertainties, provide traceability	Interactive visual explanations, model adjustments, what-if-analysis

been met, or expose (and ideally explain) failure to do so (Table 1). They mainly focus on technical aspects of CDSS along the visual data exploration loop. These should be further supported by non-technical (procedural and social) trust building measures during CDSS adoption. Non-technical measures are typically not specific to CDSS but would be advisable with the introduction of any complex system or process, so we only briefly mention aspects that pertain to CDSS in particular.

4.3.1. Technical measures

Within the scope of model-based CDSS, technical measures relate to the model component, including the data that went into building that model, and any visual analytics system component allowing users to interact with it as an actor in the joint decision process [77].

Underlying the trust in the model performance itself is trust in the underlying ground truth in terms of data quality [50,53,78] and provenance [49,79]. Relevant dimensions of data quality again depend on the specific use case, with multiple overlapping taxonomies found in literature [80], but with accuracy, privacy, fairness, generalizability, currentness, and completeness appearing to be most commonly referred. A baseline measure is to provide access to corresponding metadata for cursory data quality assessment [78]. In most real-world cases, data quality assurance and data cleansing cannot be done fully automatically due to the ambiguity of errors and the need of human knowledge to verify the cleansing results. To effectively loop humans into the data quality assurance and cleansing process as a trust-building

measure, visual analytics researchers have developed several works focusing on interactive data assessment and cleansing [77,81].

To gauge model quality, HCPs require explanation of both the model output required as well as means to interpret the models inner workings leading to a given output. At the lowest level, trust building encompasses communication of basic quality metrics, i.e., trustworthiness aspects accuracy, reliability and consistency, for relevant test data sets and task settings. Documentation of these metrics would typically utilize well-known visualizations; to design effective trust-building measures alignment with professional task requirements (e.g., thresholds, breakdown by relevant cohorts) is key [48].

However, since models will always be imperfect, situations where system output and clinicians' assessment diverge will inevitable arise. In these situations, additional explainability measures that facilitate reasoning about the divergence for HCPs to make "well-considered and trustworthy decisions" [82]. Inherently interpretable models (e.g., decision trees or rule-based models) are preferred by clinicians over post-hoc methods (e.g., SHAPley or LIME) [83]. Explainability also includes uncertainty quantification, i.e, displaying confidence scores [8]. Beyond offering evidence for a given model output, measures offering the possibility to refine algorithmic results can increase trustworthiness while also increasing diagnostic utility [8,47,84]. Providing continuous feedback can enhance clinicians' situational awareness and foster their trust in AI [60].

4.3.2. Supporting non-technical measures

External validation of results can increase a system's trustworthiness [59,61,78], as for many clinicians, validation is closely related or even a proxy to trust [7]. Clinical users need an introduction to (clinical) machine learning and AI systems [38,41,85] in general and into the specific methodology and evidence of the system in particular [46]. The mode of collaboration between human and AI should be discussed in advance, and there is a need for a calibration phase during onboarding [53]. Additionally, expectation management is needed: When people expect a system comparable to the current clinical gold standard, they quickly lose trust [49]. Expert endorsement can help clinicians build trust in AI systems, especially as they often immediately act on their colleagues' input [59,86].

In summary, individual processes to develop trust in CDSS, the set and individual weight of trustworthiness aspects depends on the individual and circumstances [14,72,76]. Thus, rather than prescribing a fixed, one-size-fits-all trust-building process, our proposed methodology aims at providing guidelines that map measures providing evidence to relevant trustworthiness factors systematically. The focus is on technical aspects and properties of a CDSS. Refer to Table 1 for the full list of proposed trustworthiness aspects and associated measures. They can be further grouped into trust quality gates along the CDSS development process, aligned with the visual data exploration loop.

4.4. Trust quality gates

We introduce *trust quality gates*, which are inspired by the namesake concept in software engineering, where certain quality criteria are defined in advance of a project [87]. Reaching and satisfying a criterion is necessary to move to the next project step. Related concepts exist in the form of a trust cue taxonomy for cognitive agents [88]. Our trust gates build upon the introduced aspects and enriches them with a way to check for sufficient trust building along the visual data exploration loop. A structured overview of the quality gates is provided in Table 1. We divide trust quality gates into the components *claim*, *requirement* and *evidence*:

- **Claims** are properties of a system component that contribute to trustworthiness. For example, a claim could be that the model provides consistent predictions across different patient demographics.

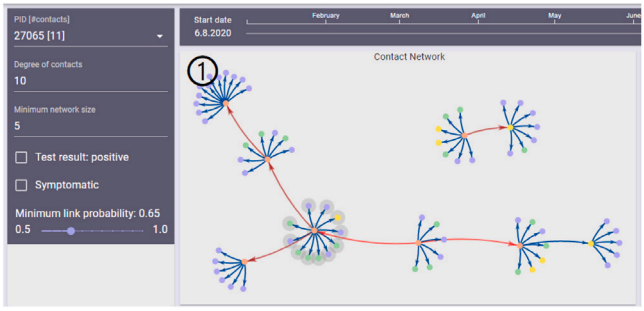

Fig. 3. *Case study example: Policy development.* A contact network shows persons with documented contacts (blue edges) as well as predicted transmissions (red edges). The user can utilize multiple filters to identify particular groups of individuals (panel to the left).

- **Requirements** are specific, application-related thresholds or standards that must be achieved concerning a claim. For instance, a requirement for the consistency claim might be a less than 5% variation in model accuracy across demographic groups.
- **Evidences** are verifiable proofs, that the requirements have been met, thereby validating the claims. Evidence can include documented results from testing phases, user feedback, or compliance with regulatory standards.

By structuring these gates around the components of claim, requirement, and evidence, we create a transparent and accountable framework for trust assessment. To operationalize these gates, the involved multidisciplinary teams collaboratively define the claims relevant to their application context. Once these claims are established, requirements can be specified, taking into account the unique challenges of the use case at hand and the environment in which the CDSS will operate. During the phases of the visual data exploration loop, the team must gather evidence to demonstrate that the requirements for each claim have been met. The third column in Table 1 highlights how visual analytics methods help during this step. Trust is therefore built progressively at each stage of development.

Our embedding of trust quality gates into the development process provides a structured methodology for systematically addressing and improving trustworthiness of ML-based CDSS. This approach not only helps to identify potential trust issues early but also ensures that trust is maintained and strengthened throughout the life cycle of the system. Benefits of this methodology include an increased transparency for stakeholders to better understand how trustworthiness is being evaluated and built into the system. Meeting regulatory requirements becomes more straightforward, supporting the safe and ethical deployment of CDSS in healthcare settings. Lastly, it leads to better communication and alignment between technical teams and healthcare professionals, assuring that trust is considered from multiple perspectives.

5. Case studies

In this section we review three case studies, each outlining a case-specific instantiation of our proposed trust methodology. During each case, we refer back to the specific trustworthiness aspects ①–⑦ from Section 4.1.

5.1. Healthcare policy development

During the COVID-19 pandemic, departments of public health (DPHs) were facing challenges to analyze the spread of infections and develop suitable containment strategies. This was mainly due to a highly dynamic number of infections, limited staff resources and

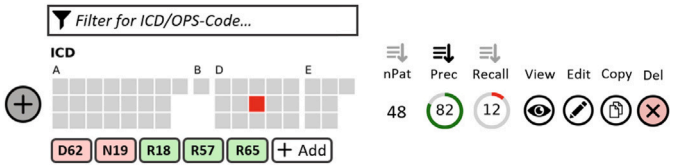

Fig. 4. *Case study example: Plausibility testing.* A list of all rules, describing their contents and metrics in detail including a rule fingerprint displaying which medical codes are inclusive (green) or exclusive (red) conditions within the rule, relevant rule metrics as well as rule editing elements to add or remove individual codes or entire rules. The rule set can be filtered and sorted to select for rules of interest.

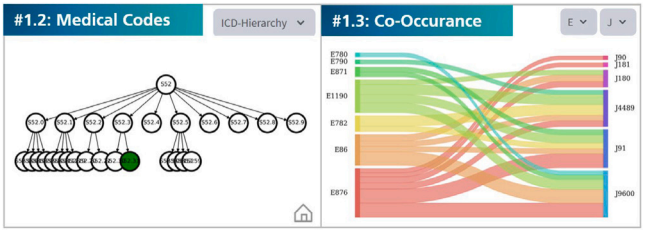

Fig. 5. *Case study example: Model optimization.* Tree visualization for medical code hierarchy next to a Sankey diagram displaying the co-occurrence of codes during patient stays.

low levels of digitization. In a joint project with Germany's largest DPH, we developed a web-based visual analytics dashboard to support healthcare analysts in identifying connected infection cases and clusters through accumulated epidemiological knowledge and metadata in their database [89].

Highly relevant trustworthiness aspects during model development included data explorability ①, generalizability to other DPHs ⑤, customizable levels of granularity ⑥, familiar digital interfaces and reduced cognitive load (cf. Table 1). To actively foster trust building, we designed the system to provide fine-granular control ④, overview and detail facets, and incorporate interfaces for data quality control. Customizable model accuracy thresholds, adherence to well-known color and naming schemes ③, and alignment with established workflows further reinforced user confidence. Explainability was achieved by employing a simple decision tree model with a small number of features ⑦, which still achieved high accuracy on our test dataset ②. Additionally, we align the visual workflow with the existing needs of public health officers by offering multiple views on different aspects of the data including geographic, temporal and social dynamics (cf. Fig. 3).

5.2. Plausibility testing

Timely detection of deteriorating conditions in hospital patients is vital for early countermeasures. At the moment, this is usually done based on clinical experience, expertise, and attention. Recently, more available data on patients and cohorts enable data-driven approaches to identify patients at risk. Together with healthcare professionals, we developed a web-based, interactive visual analytics system to assist in early detection by prediction based on collected medical codes representing diagnosis and procedures (cf. Fig. 4) [90].

Users can explore dataset selections with multiple coordinated and juxtaposed views ①. To foster trustworthiness and facilitate trust building, we prioritized high model accuracy, consistency ②, use of standardized clinical terminology, high autonomy through user control ④, flexible granularity, alignment with clinical goals, explainability, and intuitive interfaces (cf. Table 1). We achieved this by first analyzing key tasks of healthcare professionals, identifying technical requirements on a proposed solution and discussing intermediate results regularly with clinicians. Second, we used those findings as principles for our visualization methods to create a clear workflow aligned with the user's tasks ③, utilizes medical terminology ⑥, embeds model results in interactive charts and explains the model's inner workings in detail ⑦.

5.3. Model optimization

Given that patient data frequently appears as structured temporal data, many approaches for risk clinical prediction leverage Large Language Models, such as BERT. These models are notoriously complex, hard to interpret, and therefore challenging to trust. This lack of interpretability not only affects model optimization but also raises concerns regarding their trustworthiness in clinical decision-making. To address this, we created a visual analytics system dedicated to examining, comparing, and interpreting pre-trained transformer models for predicting clinical outcomes based on medical codes [91]. The use case focuses on prediction tasks for acute kidney injury (AKI) and heart failure, developed using a comprehensive hospital patient dataset.

Throughout the project, we prioritized trust building by ensuring transparency, interpretability, and usability. This included dataset exploration for bias and outlier detection ① ⑤, comfortable use by clinicians ③, alignment with clinical goals, information sufficiency, and model performance comparison and explainability ⑦. To foster trust, we implemented visual benchmark comparisons, detailed labeling and legends, individual outcome explainers, and guided analysis through a suggested examination cycle and preference options (cf. Fig. 5). Multiple views are dedicated to Accuracy, Reliability & Consistency ②, allowing users to explore different model architecture, hyperparameters, performance metrics and individual predictions. Additionally, we integrated filtered and linked views for high interactivity, hierarchical code visualizations adapted to clinical needs and visual elements to explore different training strategies. Feedback from HCPs indicates that such a system can enhance trust in model predictions, accelerate decision-making, and improve modeling results.

6. Discussion

In this work, we addressed common pitfalls during model development for CDSS concerning trust forming. First, we argue for a systematic approach to ensure trust is considered at every step of the development process, including training data, modeling, evaluation and integration. This is achieved on the one hand by examining relevant aspects from our list of identified trust clusters in Section 4.1. Additionally, we propose specific trust-building measures in Section 4.3 to support the development of trust in these aspects (*Pitfall 1*). Secondly, we differentiate between trust and trustworthiness when discussing individual aspects and their interrelation. While the subjective nature of trust is a complicated psychological and social issue, trustworthiness of digital systems can be modeled and either measured explicitly or evaluated qualitatively. Visual analytics methods can amplify trustworthiness aspects of ML models significantly by letting users explore and interact with descriptions, explanations, and views hands-on to ultimately foster trust building. (*Pitfall 2*).

We explicitly model trust as a multi-modal concept, covering a wide range of dimensions inside the seven trust clusters including privacy, accuracy, consistency, reliability, fairness, generalizability, efficiency, flexibility, ease of use, comprehensibility, explainability, autonomy, and calibration (*Pitfall 3*). Next, we consider the flow of trust between components explicitly in Section 4.2. As we already explored stakeholder relations in our original publication [9], our focus lies on highlighting how singular trust-building measures can address multiple aspects at once and accumulate along the visual data exploration loop (*Pitfall 4*). Our methodology is geared towards utilizing visual analytics methods to support trustworthy ML-based CDSS model development. By highlighting the large number of visual interfaces, representations

and views that can be employed along the process, we reveal diverse opportunities for visual analytics to augment trust building (*Pitfall 5*). We underpin our arguments by presenting case studies that showcase the diverse range of applications that benefit from trust building (Section 5).

Not surprisingly, design choices for trustworthy visualization [27] frequently overlap with trustworthiness aspects of CDSS described in Section 4.1, as healthcare professionals almost always interact with CDSS through visual interfaces. Still, healthcare professionals encounter situations of decision-making that differ profoundly from general users of visualizations described in the literature such as e-commerce, online reviews, and recommendations. Most importantly, healthcare professionals will typically assess decisions impacting a patient or the general public, not themselves. They therefore not only need to comprehend and evaluate information presented to them, but also communicate it effectively to someone with significantly lower medical literacy, on average. Both the comprehension and effective communication of visualized data must occur under significant time constraints in clinical settings. Additionally, the decisions at stake and their consequences are much more critical and healthcare professionals must therefore exercise much more caution and skepticism when interacting with a CDSS. This fact must be taken into account during CDSS development, such as placing more emphasis on incorporating uncertainty visualization or provenance tracking.

Our study proposes a general methodology rather than a specific implementation. This broad applicability allows for flexibility and adaptation, enabling a practical, real-world implementation in specific healthcare settings. The implementation of trust quality gates adds efforts during development and might meet resistance from stakeholders or technical constraints. These efforts must be weighed against potential benefits. The healthcare sector includes more key players than our case studies are covering, including outpatient care, pharmacies, and health insurers, each with unique processes, data, and use cases. We acknowledge this limitation and aim to incorporate their perspectives and expand our visual analytics methodology in future extended work accordingly.

7. Conclusion

Visual analytics is a robust toolkit for addressing the challenges of complex, data-driven, and AI-based decision-making. To bridge the gap between data scientists, visual analytics experts, and medical professionals, these systems must cultivate trust in the underlying data, models, and outcomes. We presented a novel embedding of trustworthiness aspects into the visual data exploration loop for Clinical Decision Support Systems. Our extensive literature review and subsequent systematic analysis uncovered diverse trust facets, which we consolidated into thematic groups and explored their relations in detail.

We have demonstrated that lack of trust is a significant barrier to the adoption of ML-based CDSS in healthcare. Trustworthiness must be systematically embedded throughout the development process. Visual analytics can enhance the trustworthiness of ML models by offering interactive visualizations that support both data scientists and healthcare professionals. Common pitfalls during development can be avoided by defining and assessing trust quality gates. Trust quality gates align with current and upcoming healthcare regulations (e.g., EU AI Act) and could provide a stronger argument for their adoption.

Open questions include trustworthiness regarding agentic machine learning in healthcare [92]. While current approaches foster trust through data explorability, model explainability and transparent autonomy, all these aspects may change considerably when future CDSS systems will suggests actions autonomously, request additional data or interact with available internal or external tools. Additionally, the consolidation and harmonization of trustworthiness aspects and quality gates with user-centered design in CDSS development settings is an obvious desideratum to build helpful tools for subsequent projects. For future work, we plan to focus on empirical validation of the trust quality gates introduced here across different healthcare domains and settings.

CRediT authorship contribution statement

Dario Antweiler: Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Investigation, Data curation, Conceptualization. **Georg Fuchs:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work of Antweiler was developed by the Fraunhofer Center for Machine Learning within the Fraunhofer Cluster for Cognitive Internet Technologies (CCIT).

Data availability

No data was used for the research described in the article.

References

[1] Schaffter T, Buist DSM, Lee CI, Nikulin Y, Ribli D, Guan Y, Lotter W, Jie Z, Du H, Wang S, Feng J, Feng M, Kim H-E, Albiol F, Albiol A, Morrell S, Wojna Z, Ahsen ME, Asif U, Jimeno Yepes A, Yohanandan S, Rabinovici-Cohen S, Yi D, Hoff B, Yu T, Chaibub Neto E, Rubin DL, Lindholm P, Margolies LR, McBride RB, Rothstein JH, Sieh W, Ben-Ari R, Harrer S, Trister A, Friend S, Norman T, Sahiner B, Strand F, Guinney J, Stolovitzky G, Mackey L, Cahoon J, Shen L, Sohn JH, Trivedi H, Shen Y, Buturovic L, Pereira JC, Cardoso JS, Castro E, Kalleberg KT, Pelka O, Nedjar I, Geras KJ, Nensa F, Goan E, Koitka S, Caballero L, Cox DD, Krishnaswamy P, Pandey G, Friedrich CM, Perrin D, Fookes C, Shi B, Cardoso Negrie G, Kawczynski M, Cho K, Khoo CS, Lo JY, Sorensen AG, Jung H. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. JAMA Netw Open 2020;3(3):e200265. http://dx.doi.org/10.1001/jamanetworkopen.2020.0265.

[2] Wijnberge M, Geerts BF, Hol L, Lemmers N, Mulder MP, Berge P, Schenk J, Terwindt LE, Hollmann MW, Vlaar AP, Veelo DP. Effect of a machine learning–derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: The HYPE randomized clinical trial. JAMA 2020;323(11):1052. http://dx.doi.org/10.1001/jama.2020.0592.

[3] Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nature Med 2018;24(11):1716–20. http://dx.doi.org/10.1038/s41591-018-0213-5.

[4] Kugler S, Antweiler D, Stein B. From rehab to recap: Generating discharge reports for rehabilitation clinics with LLMs. In: 2024 2nd international conference on foundation and large language models. FLLM, IEEE; 2024, p. 378–84. http://dx.doi.org/10.1109/fllm63129.2024.10852496.

[5] Kamel Rahimi A, Ghadimi M, van der Vegt AH, Canfell OJ, Pole JD, Sullivan C, Shrapnel S. Machine learning clinical prediction models for acute kidney injury: the impact of baseline creatinine on prediction efficacy. BMC Med Inform Decis Mak 2023;23(1). http://dx.doi.org/10.1186/s12911-023-02306-0.

[6] Colantonio S, Berti A, Buongiorno R, Del Corso G, Pachetti E, Pascali MA, Kalantzopoulos C, Kalokyri V, Kondylakis H, Tachos N, Fotiadis D, Giannini V, Mazzetti S, Regge D, Papanikolaou N, Marias K, Tsiknakis M. AI trustworthiness in prostate cancer imaging: a look at algorithmic and system transparency *. In: 2023 IEEE EMBS special topic conference on data science and engineering in healthcare, medicine and biology. Malta: IEEE; 2023, p. 79–80. http://dx.doi.org/10.1109/IEEECONF58974.2023.10404432.

[7] Winter P, Carusi A. 'If you're going to trust the machine, then that trust has got to be based on something': Validation and the co-constitution of trust in developing artificial intelligence (AI) for the early diagnosis of pulmonary hypertension (PH). Sci Technol Stud 2022. http://dx.doi.org/10.23987/sts.102198.

[8] Cutillo CM, Sharma KR, Foschini L, Kundu S, Mackintosh M, Mandl KD, Beck T, Collier E, Colvis C, Gersing K, Gordon V, Jensen R, Shabestari B, Southall N. Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. Npj Digit Med 2020;3(1). http://dx.doi.org/10.1038/s41746-020-0254-2.

[9] Antweiler D, Fuchs G. Extending the visual data exploration loop towards trustworthy machine learning in the healthcare domain. In: El-Assady M, Schulz H-J, editors. EuroVis workshop on visual analytics (euroVA). The Eurographics Association; 2024, http://dx.doi.org/10.2312/eurova.20241107.

[10] Keim DA, Mansmann F, Stoffel A, Ziegler H. Visual analytics. In: Encyclopedia of database systems. Springer US; 2009, p. 3341–6. http://dx.doi.org/10.1007/978-0-387-39940-9_1122.

[11] O'Neill O. Linking trust to trustworthiness. Int J Philos Stud 2018;26(2):293–300. http://dx.doi.org/10.1080/09672559.2018.1454637.

[12] Cho J-H, Chan K, Adali S. A survey on trust modeling. ACM Comput Surv 2015;48(2):1–40. http://dx.doi.org/10.1145/2815595.

[13] Varshney KR. Trustworthy machine learning. 2022.

[14] Felici M. How to trust: A model for trust decision making. Int J Adapt Resilient Auton Syst 2012;3(3):20–34. http://dx.doi.org/10.4018/jaras.2012070102.

[15] Mcknight D. Trust in information technology. The Blackwell Encyclopedia of Management; 2005, p. 329–31.

[16] Hasani N, Morris MA, Rahmim A, Summers RM, Jones E, Siegel E, Saboury B. Trustworthy artificial intelligence in medical imaging. PET Clin 2022;17(1):1–12. http://dx.doi.org/10.1016/j.cpet.2021.09.007.

[17] Chatzimparmpas A, Martins RM, Jusufi I, Kucher K, Rossi F, Kerren A. The state of the art in enhancing trust in machine learning models with the use of visualizations. Comput Graph Forum 2020;39(3):713–56. http://dx.doi.org/10.1111/cgf.14034.

[18] Elzen Svd, Andrienko G, Andrienko N, Fisher BD, Martins RM, Peltonen J, Telea AC, Verleysen M. The flow of trust: A visualization framework to externalize, explore, and explain trust in ML applications. IEEE Comput Graph Appl 2023;43(2):78–88. http://dx.doi.org/10.1109/mcg.2023.3237286.

[19] Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: Focus on clinicians. J Med Internet Res 2020;22(6):e15154. http://dx.doi.org/10.2196/15154.

[20] Kinkeldey C, Korjakow T, Benjamin JJ. Towards supporting interpretability of clustering results with uncertainty visualization. Trust EuroVis 2019. http://dx.doi.org/10.2312/TRVIS.20191183.

[21] Ragan ED, Endert A, Sanyal J, Chen J. Characterizing provenance in visualization and data analysis: An organizational framework of provenance types and purposes. IEEE Trans Vis Comput Graph 2016;22(1):31–40. http://dx.doi.org/10.1109/tvcg.2015.2467551.

[22] Yuan J, Barr B, Overton K, Bertini E. Visual exploration of machine learning model behavior with hierarchical surrogate rule sets. 2022, http://dx.doi.org/10.48550/arxiv.2201.07724.

[23] Chatzimparmpas A, Martins RM, Telea AC, Kerren A. DeforestVis: Behavior analysis of machine learning models with surrogate decision stumps. 2023, http://dx.doi.org/10.48550/arxiv.2304.00133.

[24] Dasgupta A, Lee J-Y, Wilson R, Lafrance RA, Cramer N, Cook K, Payne S. Familiarity vs Trust: A comparative study of domain scientists' trust in visual analytics and conventional analysis methods. IEEE TVCG 2017;23(1):271–80. http://dx.doi.org/10.1109/tvcg.2016.2598544.

[25] Beauxis-Aussalet E, Behrisch M, Borgo R, Chau DH, Collins C, Ebert D, El-Assady M, Endert A, Keim DA, Kohlhammer J, Oelke D, Peltonen J, Riveiro M, Schreck T, Strobelt H, van Wijk JJ. The role of interactive visualization in fostering trust in AI. IEEE Comput Graph Appl 2021;41(6):7–12. http://dx.doi.org/10.1109/MCG.2021.3107875.

[26] Jia Y, McDermid J, Lawton T, Habli I. The role of explainability in assuring safety of machine learning in healthcare. IEEE TETC 2022;10(4):1746–60. http://dx.doi.org/10.1109/tetc.2022.3171314.

[27] Wall E, Matzen L, El-Assady M, Masters P, Hosseinpour H, Endert A, Borgo R, Chau P, Perer A, Schupp H, Strobelt H, Padilla L. Trust junk and evil knobs: Calibrating trust in AI visualization. In: 2024 IEEE 17th Pacific visualization conference (Pacificvis). IEEE; 2024, p. 22–31. http://dx.doi.org/10.1109/pacificvis60374.2024.00012.

[28] Fliegenschmidt J, Hulde N, Preising MG, Ruggeri S, Szymanowski R, Meesseman L, Sun H, von Dossow V. Artificial intelligence predicts delirium following cardiac surgery: A case study. J Clin Anesth 2021;75:110473. http://dx.doi.org/10.1016/j.jclinane.2021.110473.

[29] Sendak MP, Ratliff W, Sarro D, Alderton E, Futoma J, Gao M, Nichols M, Revoir M, Yashar F, Miller C, Kester K, Sandhu S, Corey K, Brajer N, Tan C, Lin A, Brown T, Engelbosch S, Anstrom K, Elish MC, Heller K, Donohoe R, Theiling J, Poon E, Balu S, Bedoya A, O'Brien C. Real-world integration of a sepsis deep learning technology into routine clinical care: Implementation study. JMIR Med Inform 2020;8(7):e15182. http://dx.doi.org/10.2196/15182.

[30] Wang X, Sontag D, Wang F. Unsupervised learning of disease progression models. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining. ACM; 2014, http://dx.doi.org/10.1145/2623330.2623754.

[31] Bihorac A, Ozrazgat-Baslanti T, Ebadi A, Motaei A, Madkour M, Pardalos PM, Lipori G, Hogan WR, Efron PA, Moore F, Moldawer LL, Wang DZ, Hobson CE, Rashidi P, Li X, Momcilovic P. MySurgeryRisk: Development and validation of a machine-learning risk algorithm for major complications and death after surgery. Ann Surg 2019;269(4):652–62. http://dx.doi.org/10.1097/sla.0000000000002706.

[32] Shoeb A, Guttag J. Application of machine learning to epileptic seizure detection. In: Proceedings of the 27th international conference on international conference on machine learning. ICML '10, Madison, WI, USA: Omni Press; 2010, p. 975–82.

[33] Sorin V, Klang E, Sklair-Levy M, Cohen I, Zippel DB, Lahat NB, Konen E, Barash Y. Large language model (ChatGPT) as a support tool for breast tumor board. Npj Breast Cancer 2023;9(1). http://dx.doi.org/10.1038/s41523-023-00557-8.

[34] Kwan JL, Lo L, Ferguson J, Goldberg H, Diaz-Martinez JP, Tomlinson G, Grimshaw JM, Shojania KG. Computerised clinical decision support systems and absolute improvements in care: meta-analysis of controlled clinical trials. BMJ 2020;m3216. http://dx.doi.org/10.1136/bmj.m3216.

[35] Schütze D, Holtz S, Neff MC, Köhler SM, Schaaf J, Frischen LS, Sedlmayr B, Müller BS. Requirements analysis for an AI-based clinical decision support system for general practitioners: a user-centered design process. BMC Med Inform Decis Mak 2023;23(1). http://dx.doi.org/10.1186/s12911-023-02245-w.

[36] Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. Npj Digit Med 2020;3(1). http://dx.doi.org/10.1038/s41746-020-0221-y.

[37] Ash JS, Sittig DF, Campbell EM, Guappone KP, Dykstra RH. Some unintended consequences of clinical decision support systems. In: AMIA Annu Symp Proc. 2007, p. 26–30., URL https://pmc.ncbi.nlm.nih.gov/articles/PMC2813668/.

[38] Zhang J, Zhang Z-m. Ethics and governance of trustworthy medical artificial intelligence. BMC Med Inform Decis Mak 2023;23(1). http://dx.doi.org/10.1186/s12911-023-02103-9.

[39] Schwabe D, Becker K, Seyferth M, Klaß A, Schaeffter T. The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. Npj Digit Med 2024;7(1). http://dx.doi.org/10.1038/s41746-024-01196-4.

[40] Chew HSJ, Achananuparp P. Perceptions and needs of artificial intelligence in health care to increase adoption: Scoping review. J Med Internet Res 2022;24(1):e32939. http://dx.doi.org/10.2196/32939.

[41] Sandhu S, Lin AL, Brajer N, Sperling J, Ratliff W, Bedoya AD, Balu S, O'Brien C, Sendak MP. Integrating a machine learning system into clinical workflows: Qualitative study. J Med Internet Res 2020;22(11):e22421. http://dx.doi.org/10.2196/22421.

[42] Zuo Z, Watson M, Budgen D, Hall R, Kennelly C, Al Moubayed N. Data anonymization for pervasive health care: Systematic literature mapping study. JMIR Med Inform 2021;9(10):e29871. http://dx.doi.org/10.2196/29871.

[43] Carlini N, Tramèr F, Wallace E, Jagielski M, Herbert-Voss A, Lee K, Roberts A, Brown T, Song D, Erlingsson U, Oprea A, Raffel C. Extracting training data from large language models. In: 30th USeNIX security symposium (USeNIX security 21). USENIX Association; 2021, p. 2633–50.

[44] Esmaeilzadeh P, Mirzaei T, Dharanikota S. Patients' perceptions toward human–artificial intelligence interaction in health care: Experimental study. J Med Internet Res 2021;23(11):e25856. http://dx.doi.org/10.2196/25856.

[45] Rossi A, Lenzini G. Making the case for evidence-based standardization of data privacy and data protection visual indicators. J Open Access L 2020;8:1.

[46] Sivaraman V, Bukowski LA, Levin J, Kahn JM, Perer A. Ignore, trust, or negotiate: Understanding clinician acceptance of AI-based treatment recommendations in health care. In: Proceedings of the 2023 CHI conference on human factors in computing systems. CHI '23, ACM; 2023, p. 1–18. http://dx.doi.org/10.1145/3544548.3581075.

[47] Li W, Fan X, Zhu H, Wu J, Teng D. Research on the influencing factors of user trust based on artificial intelligence self diagnosis system. In: Proceedings of the ACM turing celebration conference - China. Hefei China: ACM; 2020, p. 197–202. http://dx.doi.org/10.1145/3393527.3393561.

[48] Schwartz JM, George M, Rossetti SC, Dykes PC, Minshall SR, Lucas E, Cato KD. Factors influencing clinician trust in predictive clinical decision support systems for in-hospital deterioration: Qualitative descriptive study. JMIR Hum Factors 2022;9(2):e33960. http://dx.doi.org/10.2196/33960.

[49] Joshi M, Mecklai K, Rozenblum R, Samal L. Implementation approaches and barriers for rule-based and machine learning-based sepsis risk prediction tools: a qualitative study. JAMIA Open 2022;5(2). http://dx.doi.org/10.1093/jamiaopen/ooac022.

[50] Shibl R, Lawley M, Debuse J. Factors influencing decision support system acceptance. Decis Support Syst 2013;54(2):953–61. http://dx.doi.org/10.1016/j.dss.2012.09.018.

[51] Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach Learn 2021;110(3):457–506. http://dx.doi.org/10.1007/s10994-021-05946-3.

[52] Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What clinicians want: Contextualizing explainable machine learning for clinical end use. 2019, http://dx.doi.org/10.48550/arxiv.1905.05134.

[53] Cai CJ, Winter S, Steiner D, Wilcox L, Terry M. "Hello AI": Uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. Proc ACM Hum- Comput Interact 2019;3(CSCW). http://dx.doi.org/10.1145/3359206.

[54] Zytek A, Liu D, Vaithianathan R, Veeramachaneni K. Sibyl: Understanding and addressing the usability challenges of machine learning in high-stakes decision making. 2021, http://dx.doi.org/10.48550/ARXIV.2103.02071, URL https://arxiv.org/abs/2103.02071.

[55] Berti A, Buongiorno R, Carloni G, Caudai C, Corso GD, Germanese D, Pachetti E, Pascali MA, Colantonio S. Exploring the potentials and challenges of Artificial Intelligence in supporting clinical diagnostics and remote assistance for the health and well-being of individuals. In: Falchi F, Giannotti F, Monreale A, Boldrini C, Rinzivillo S, Colantonio S, editors. Proceedings of the Italia intelligenza artificiale - thematic workshops co-located with the 3rd CINI national lab AIIS conference on artificial intelligence (ital IA 2023), pisa, Italy, May 29-30, 2023. CEUR workshop proceedings, Vol. 3486, CEUR-WS.org; 2023, p. 146–53, URL https://ceur-ws.org/Vol-3486/110.pdf.

[56] Cabitza F, Campagner A, Balsano C. Bridging the "last mile" gap between AI implementation and operation: "data awareness" that matters. Ann Transl Med 2020;8(7):501. http://dx.doi.org/10.21037/atm.2020.03.63.

[57] Scott IA, Carter SM, Coiera E. Exploring stakeholder attitudes towards AI in clinical practice. BMJ Heal Care Inform 2021;28(1):e100450. http://dx.doi.org/10.1136/bmjhci-2021-100450.

[58] Feldman R, Aldana E, Stein K. Artificial intelligence in the health care space: How we can trust what we cannot know. Stanf Law Policy Rev 2019;30:399, URL https://repository.uclawsf.edu/faculty_scholarship/1753, Available at U.C. Law SF Repository.

[59] Henry KE, Kornfield R, Sridharan A, Linton RC, Groh C, Wang T, Wu A, Mutlu B, Saria S. Human–machine teaming is key to AI adoption: clinicians' experiences with a deployed machine learning system. Npj Digit Med 2022;5(1). http://dx.doi.org/10.1038/s41746-022-00597-7.

[60] Sujan MA, White S, Habli I, Reynolds N. Stakeholder perceptions of the safety and assurance of artificial intelligence in healthcare. Saf Sci 2022;155:105870. http://dx.doi.org/10.1016/j.ssci.2022.105870.

[61] Jacobs M, He J, F. Pradier M, Lam B, Ahn AC, McCoy TH, Perlis RH, Doshi-Velez F, Gajos KZ. Designing AI for trust and collaboration in time-constrained medical decisions: A sociotechnical lens. In: Proceedings of the 2021 CHI conference on human factors in computing systems. Yokohama Japan: ACM; 2021, p. 1–14. http://dx.doi.org/10.1145/3411764.3445385.

[62] Diprose WK, Buist N, Hua N, Thurier Q, Shand G, Robinson R. Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. J Am Med Inform Assoc 2020;27(4):592–600. http://dx.doi.org/10.1093/jamia/ocz229.

[63] Fan W, Liu J, Zhu S, Pardalos PM. Investigating the impacting factors for the healthcare professionals to adopt artificial intelligence-based medical diagnosis support system (AIMDSS). Ann Oper Res 2020;294(1–2):567–92. http://dx.doi.org/10.1007/s10479-018-2818-y.

[64] Dey S, Chakraborty P, Kwon BC, Dhurandhar A, Ghalwash M, Suarez Saiz FJ, Ng K, Sow D, Varshney KR, Meyer P. Human-centered explainability for life sciences, healthcare, and medical informatics. Patterns 2022;3(5):100493. http://dx.doi.org/10.1016/j.patter.2022.100493.

[65] Han W, Schulz H-J. Beyond trust building — Calibrating trust in visual analytics. In: 2020 IEEE workshop on tRust and eXpertise in visual analytics. TREX, 2020, p. 9–15. http://dx.doi.org/10.1109/TREX51495.2020.00006.

[66] Micocci M, Borsci S, Thakerar V, Walne S, Manshadi Y, Edridge F, Mullarkey D, Buckle P, Hanna GB. Attitudes towards trusting artificial intelligence insights and factors to prevent the passive adherence of GPs: A pilot study. J Clin Med 2021;10(14):3101. http://dx.doi.org/10.3390/jcm10143101.

[67] van der Vegt AH, Scott IA, Dermawan K, Schnetler RJ, Kalke VR, Lane PJ. Implementation frameworks for end-to-end clinical AI: derivation of the SALIENT framework. J Am Med Inform Assoc 2023;30(9):1503–15. http://dx.doi.org/10.1093/jamia/ocad088.

[68] Farah L, Murris JM, Borget I, Guilloux A, Martelli NM, Katsahian SI. Assessment of performance, interpretability, and explainability in artificial intelligence–based health technologies: What healthcare stakeholders need to know. Mayo Clin Proc: Digit Heal 2023;1(2):120–38. http://dx.doi.org/10.1016/j.mcpdig.2023.02.004.

[69] Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, Liu PJ, Liu X, Marcus J, Sun M, Sundberg P, Yee H, Zhang K, Zhang Y, Flores G, Duggan GE, Irvine J, Le Q, Litsch K, Mossin A, Tansuwan J, Wang D, Wexler J, Wilson J, Ludwig D, Volchenboum SL, Chou K, Pearson M, Madabushi S, Shah NH, Butte AJ, Howell MD, Cui C, Corrado GS, Dean J. Scalable and accurate deep learning with electronic health records. Npj Digit Med 2018;1(1). http://dx.doi.org/10.1038/s41746-018-0029-1.

[70] Bayer S, Gimpel H, Markgraf M. The role of domain expertise in trusting and following explainable AI decision support systems. J Decis Syst 2021;32(1):110–38. http://dx.doi.org/10.1080/12460125.2021.1958505.

[71] Truong T, Gilbank P, Johnson-Cover K, Ieraci A. A framework for applied AI in healthcare. Stud Health Technol Inform 2019;264:1993–4. http://dx.doi.org/10.3233/SHTI190751.

[72] Wang Y, Xu X, Jin T, Li X, Xie G, Wang J. Inpatient2Vec: Medical representation learning for inpatients. In: 2019 IEEE international conference on bioinformatics and biomedicine. BIBM, IEEE; 2019, http://dx.doi.org/10.1109/bibm47256.2019.8983281.

[73] Terry AL, Kueper JK, Beleno R, Brown JB, Cejic S, Dang J, Leger D, McKay S, Meredith L, Pinto AD, Ryan BL, Stewart M, Zwarenstein M, Lizotte DJ. Is primary health care ready for artificial intelligence? What do primary health care stakeholders say? BMC Med Inform Decis Mak 2022;22(1):237. http://dx.doi.org/10.1186/s12911-022-01984-6.

[74] Esmaeilzadeh P. Use of AI-based tools for healthcare purposes: a survey study from consumers' perspectives. BMC Med Inform Decis Mak 2020;20(1):170. http://dx.doi.org/10.1186/s12911-020-01191-1.

[75] Bussone A, Stumpf S, O'Sullivan D. The role of explanations on trust and reliance in clinical decision support systems. In: 2015 international conference on healthcare informatics. 2015, p. 160–9. http://dx.doi.org/10.1109/ICHI.2015.26.

[76] Shane German E, Rhodes DH. Model-centric decision-making: Exploring decision-maker trust and perception of models. In: Madni AM, Boehm B, Ghanem RG, Erwin D, Wheaton MJ, editors. Disciplinary convergence in systems engineering research. Cham: Springer International Publishing; 2018, p. 813–27.

[77] Liu S, Andrienko G, Wu Y, Cao N, Jiang L, Shi C, Wang Y-S, Hong S. Steering data quality with visual analytics: The complexity challenge. Vis Inform 2018;2(4):191–7. http://dx.doi.org/10.1016/j.visinf.2018.12.001, URL https://www.sciencedirect.com/science/article/pii/S2468502X18300573.

[78] Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. J Biomed Inform 2021;113:103655. http://dx.doi.org/10.1016/j.jbi.2020.103655.

[79] Lugtenberg M, Weenink J-W, Van Der Weijden T, Westert GP, Kool RB. Implementation of multiple-domain covering computerized decision support systems in primary care: a focus group study on perceived barriers. BMC Med Inform Decis Mak 2015;15(1):82. http://dx.doi.org/10.1186/s12911-015-0205-z.

[80] Sidi F, Shariat Panahy PH, Affendey LS, Jabar MA, Ibrahim H, Mustapha A. Data quality: A survey of data quality dimensions. In: 2012 international conference on information retrieval & knowledge management. 2012, p. 300–4. http://dx.doi.org/10.1109/InfRKM.2012.6204995.

[81] Van den Broeck J, Argeseanu Cunningham S, Eeckels R, Herbst K. Data cleaning: Detecting, diagnosing, and editing data abnormalities. PLOS Med 2005;2(10). http://dx.doi.org/10.1371/journal.pmed.0020267.

[82] Ooge J, Stiglic G, Verbert K. Explaining artificial intelligence with visual analytics in healthcare. WIREs Data Min Knowl Discov 2021;12(1). http://dx.doi.org/10.1002/widm.1427.

[83] Khare SK, Acharya UR. Adazd-Net: Automated adaptive and explainable Alzheimer's disease detection system using EEG signals. Knowl-Based Syst 2023;278:110858. http://dx.doi.org/10.1016/j.knosys.2023.110858.

[84] Cai CJ, Reif E, Hegde N, Hipp J, Kim B, Smilkov D, Wattenberg M, Viegas F, Corrado GS, Stumpe MC, Terry M. Human-centered tools for coping with imperfect algorithms during medical decision-making. In: Proceedings of the 2019 CHI conference on human factors in computing systems. CHI '19, ACM; 2019, http://dx.doi.org/10.1145/3290605.3300234.

[85] Strohm L, Hehakaya C, Ranschaert ER, Boon WPC, Moors EHM. Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. Eur Radiol 2020;30(10):5525–32. http://dx.doi.org/10.1007/s00330-020-06946-y.

[86] Yang Q, Zimmerman J, Steinfeld A, Carey L, Antaki JF. Investigating the heart pump implant decision process: Opportunities for decision support tools to help. In: Proceedings of the 2016 CHI conference on human factors in computing systems. San Jose California USA: ACM; 2016, p. 4477–88. http://dx.doi.org/10.1145/2858036.2858373, URL https://dl.acm.org/doi/10.1145/2858036.2858373.

[87] Ambartsoumian V, Dhaliwal J, Lee E, Meservy T, Zhang C. Implementing quality gates throughout the enterprise it production process. J Inf Technol Manag 2011;22(1):28–38.

[88] de Visser EJ, Cohen M, Freedy A, Parasuraman R. A design methodology for trust cue calibration in cognitive agents. In: Virtual, augmented and mixed reality. designing and developing virtual and augmented environments. Springer International Publishing; 2014, p. 251–62. http://dx.doi.org/10.1007/978-3-319-07458-0_24.

[89] Antweiler D, Sessler D, Rossknecht M, Abb B, Ginzel S, Kohlhammer J. Uncovering chains of infections through spatio-temporal and visual analysis of COVID-19 contact traces. C G 2022;106:1–8. http://dx.doi.org/10.1016/j.cag.2022.05.013.

[90] Antweiler D, Fuchs G. Visualizing rule-based classifiers for clinical risk prognosis. In: 2022 IEEE (VIS). IEEE; 2022, http://dx.doi.org/10.1109/vis54862.2022.00020.

[91] Antweiler D, Fuchs G, Gallusser F. Multi-task transformer visualization to build trust for clinical outcome prediction. In: VAHC 2023 (VIS). IEEE; 2023.

[92] Li J, Wang S, Zhang M, Li W, Lai Y, Kang X, Ma W, Liu Y. Agent hospital: A simulacrum of hospital with evolvable medical agents. 2024, http://dx.doi.org/10.48550/ARXIV.2405.02957.