Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202012.0313.v2

Peer-reviewed version available at *Symmetry* **2021**, *13*, 102; doi:10.3390/sym13010102

Type of the Paper (Article)

Controlling Safety of Artificial Intelligence-Based Systems in Healthcare

Mohammad Reza Davahli[1*], Waldemar Karwowski[1], Krzysztof Fiok[1], Thomas Wan[2], and Hamid R. Parsaei[3]

[1]Department of Industrial Engineering and Management Systems, University of Central Florida, Orlando, FL 32816, USA
[2]Health Management and Informatics, University of Central Florida, Orlando, FL 32816, USA
[3]Department of Industrial & Systems Engineering, Texas A&M University, College Station, TX, USA
*Correspondence: Mohammad Reza Davahli
E-mails: mohammadreza.davahli@ucf.edu (M.R. Davahli), wkar@ucf.edu (W. Karwowski), fiok@ucf.edu (K. Fiok) Thomas.Wan@ucf.edu (T. Wan) and hamid.parsaei@tamu.edu (H. R. Parsaei)

Abstract: In response to the need to address the safety challenges in the use of artificial intelligence (AI), this research aimed to develop a framework for a safety controlling system (SCS) to address the AI black-box mystery in the healthcare industry. The main objective was to propose safety guidelines for implementing AI black-box models to reduce the risk of potential healthcare-related incidents and accidents. The system was developed by adopting the multi-attribute value model approach (MAVT), which comprises four symmetrical parts: extracting attributes, generating weights for the attributes, developing a rating scale, and finalizing the system. On the basis of the MAVT approach, three layers of attributes were created. The first level contained 6 key dimensions, the second level included 14 attributes, and the third level comprised 78 attributes. The key first level dimensions of the SCS included safety policies, incentives for clinicians, clinician and patient training, communication and interaction, planning of actions, and control of such actions. The proposed system may provide a basis for detecting AI utilization risks, preventing incidents from occurring, and developing emergency plans for AI-related risks. This approach could also guide and control the implementation of AI systems in the healthcare industry.

Keywords: artificial intelligence, human-AI interaction, human factors, safety challenges, black-box challenge

© 2021 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202012.0313.v2

Peer-reviewed version available at Symmetry 2021, 13, 102; doi:10.3390/sym13010102

1. Introduction

Artificial intelligence (AI) has recently experienced substantial growth across different industries, including medicine and healthcare [1,2]. Applications of AI in healthcare can be divided into four symmetrical periods: the beginning (1956–1970), the first generation of AI algorithms in healthcare (1970–2012), the second generation of AI algorithms in healthcare (2012–2016), and AI outperforming its human counterparts in some medical tasks (2016 to the present).

The beginning. The field of AI began in 1956 as the science and engineering of making intelligent machines [3]. The first AI model, which had limited problem-solving ability, was developed in the mid-1950s [4]. In 1959, the term "machine learning" was coined by Arthur Samuel, who defined it as the field of study that gives computers the ability to learn without being programmed [5]. In 1961, an AI model was trained with data from 1035 patients and used to diagnose congenital heart disease [6]. In 1966, a natural language processing program was developed to mimic human therapists [7].

The first generation of AI algorithms in healthcare. In 1972, the performance of computer-aided diagnosis for acute abdominal pain was compared with that of human physicians. The system's overall diagnostic accuracy was higher than that of the physicians [8].

Subsequently, rule-based approaches achieved many advances in diagnosing diseases [9]. These approaches heavily relied on developing robust decision rules and using expert knowledge in medical practice [9]. The period included other important milestones. For example, in 1991, a pioneering attempt to open the black-box was conducted by Dean Pomerleau [10]. In 2004, adversarial inputs were formally described as intentionally designed input data to force AI systems to make mistakes [11].

The second generation of AI algorithms in healthcare. This period started with the development of a deep neural network-based system able to recognize a cat in pictures and videos [12]. Subsequently, deep learning attracted the attention of many researchers [12,13]. The second generation AI algorithms, in contrast to the rule-based approach, were able to analyze complex interactions in health data and discover hidden patterns [14].

AI outperforming its human counterparts in some medical tasks. Since 2016, the application of AI in healthcare has achieved considerable success, and AI models have accomplished various medical subtasks at or above the performance levels of physicians [15]. A highly accurate neural network algorithm was developed in ophthalmology for detecting diabetic retinopathy after training with manually labeled retinal fundus photographs [16]. In radiology, a convolutional neural network trained with labeled frontal chest X-ray images outperformed radiologists in detecting pneumonia [15,17]. In cardiology, a deep learning algorithm diagnosed heart attack with a performance comparable to that of (human) cardiologists [18]. In pathology, one study trained AI algorithms with whole-slide pathology images to detect lymph node metastases of breast cancer and compared the results with those of pathologists [19]. In dermatology, a convolutional neural network was trained with clinical images and was found to classify skin lesions accurately [20].

Despite these advancements, various challenges exist in applying AI in healthcare [1,2,21]. One of the main challenges is safety. Several reports have described unsafe and incorrect recommendations by AI algorithms [22]. The safety of AI models is mainly associated with model interpretability and explainability [1]. Interpretability is defined as the ability to understand how an AI model reaches its decisions [1].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202012.0313.v2

Peer-reviewed version available at Symmetry 2021, 13, 102; doi:10.3390/sym13010102

Regarding interpretability, AI models can be categorized into white-box models, such as decision trees, and black-box models, such as neural networks [23]. Compared with white-box models, black-box models have excellent performance with almost no interpretability [24].

To address the AI black-box challenge, a considerable amount of research has focused on developing explainable AI to open the black-box [23]. As a primary method for addressing the AI black-box issue, the visualization approach was developed to explain the models' main features [25]. For example, De Fauw et al. [26] visualized sections of the patient optical coherence tomography scans used by an AI model to make medical decisions. However, visualization is challenging to explain, and users tend to misread the results and over-trust their judgement [27]. Other approaches for addressing the AI black-box issue have been developed, such as (1) analyzing one isolated layer at a time to learn the differences between layers in neural networks [28], (2) using a simplified version of the algorithm for debugging and detecting potential errors, and then training an accurate version of the algorithm [29], (3) training the black-box model to explain the level of safety by assigning a confidence level to the model's prediction [30]. However, these methods focus on diminishing the black-box rather than opening the black-box of AI [26]. To open the black-box, the logic behind AI models' decision-making processes must be identified, and specific model tasks must be able to be paused or modified as necessary [31].

In contrast, some researchers are less concerned about opening the black-box of AI [28]. From this standpoint, understanding how an AI model makes decisions is less crucial than empirically verifying its accuracy [32]. According to this viewpoint, regulators and clinicians should accept the AI black-box models, because opaque systems are common in medicine [10]. For example, several efficient medications such as aspirin and penicillin were used before their mechanisms were discovered [33]. Because of the excellent performance and popularity of AI black-box models, and given the absence of effective methods to open the black-box, accepting AI black-box models could be considered an acceptable option. However, addressing the safety issues of AI black-box models is also essential [33–35].

The present study focused on developing a tool to evaluate the safety practices of AI models implemented in healthcare. The main objective of this article was to build safety guidelines for implemented AI black-box models to reduce the risk of health-related incidents and accidents. For this purpose, a three-level multi-attribute value model (MAVT) approach was used to develop a safety controlling system (SCS) for AI systems implementation.

2. Methodology

The SCS for AI implementation was developed by using a three-level MAVT adapted from Teo and Ling [36]. This approach consisted of four parts: (1) extracting attributes at different levels; (2) generating weights for the attributes; (3) assigning a rating scale for the attributes; and (4) finalizing the system [36] (see Figure 1). Several techniques were used to accomplish these steps. A combination of a systematic literature review and expert interviews was used for extracting attributes; a questionnaire-based survey was used for generating weights; and questionnaire-based survey was used for developing a rating scale.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202012.0313.v2

Peer-reviewed version available at Symmetry 2021, 13, 102; doi:10.3390/sym13010102

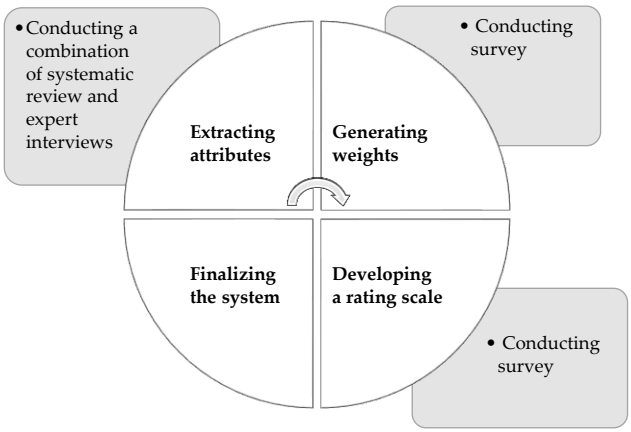

Figure 1. Components of the three-level multi-attribute value approach.

2.1. List of attributes

In the proposed SCS, the attributes were divided into three levels. The first level attributes, called key dimensions, were adopted from Fernández-Muñiz et al. [37]. These were extracted from applicable safety standards and guidelines. These key dimensions were the fundamental and well-known elements of any robust safety management system and included safety policies, incentives for clinicians, clinician and patient training, communication and interaction, the planning of actions, and the control of actions.

The second and third level attributes were developed by using a systematic literature review and interviewing ten AI domain experts. As the lowest level, the third level attributes were measurable safety elements for implemented AI systems in healthcare. The third level attributes were extracted from the systematic literature review and were subsequently refined during expert interviews. The third level attributes were clustered according to their predominant topics. These topics were named as the second level attributes.

The main reasons for using a combination of systematic review and interviews was to identify the main topics of AI safety in the included literature and expand these topics through consultation with ten AI domain experts. In addition, we aimed to ensure that all main aspects (elements) of AI implementation safety were addressed. For this purpose, first, a systematic review was conducted, and the main elements of safety in different key dimensions were extracted. Second, the extracted information was categorized and discussed during interviews with AI domain experts to produce the third attributes, as illustrated in Figure 2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202012.0313.v2

Peer-reviewed version available at Symmetry 2021, 13, 102; doi:10.3390/sym13010102

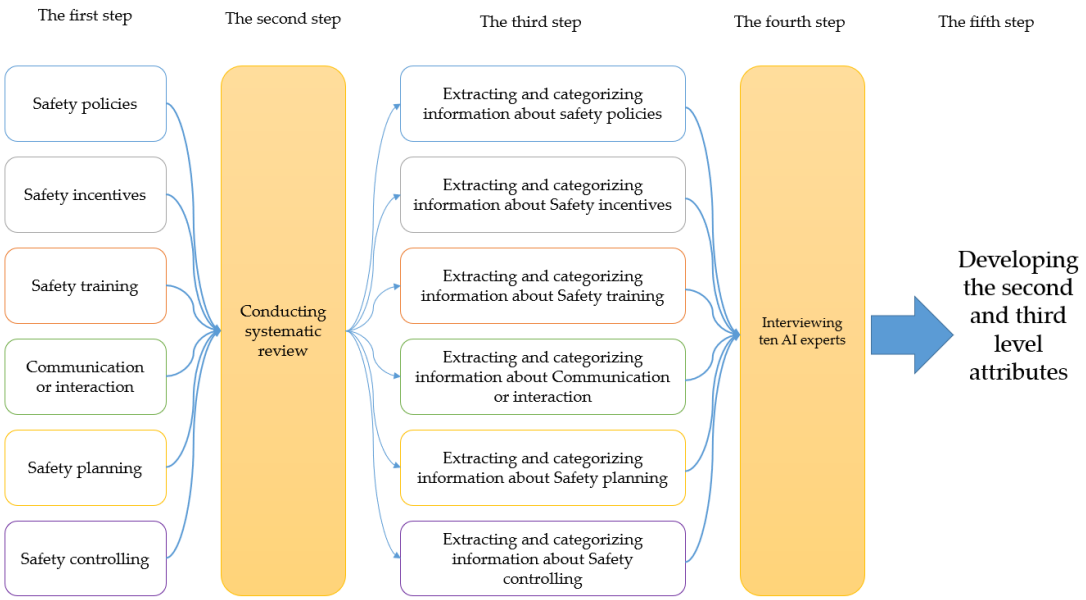

Figure 2. Approach for a combination of systematic review and interviews.

2.1.1. Systematic review

To identify the attributes, we followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [38]. Two main features of the research question and search strategy were developed. The following research question guided the systematic review:

SRQ: What are the primary safety attributes of implemented AI models in healthcare for each key dimension?

A search strategy was developed by (1) defining keywords and identifying all relevant records, (2) filtering the identified articles, and (3) addressing the risk of bias among records[39]. Three sets of keywords were defined, and their combinations were used to identify relevant articles (Figure 3).

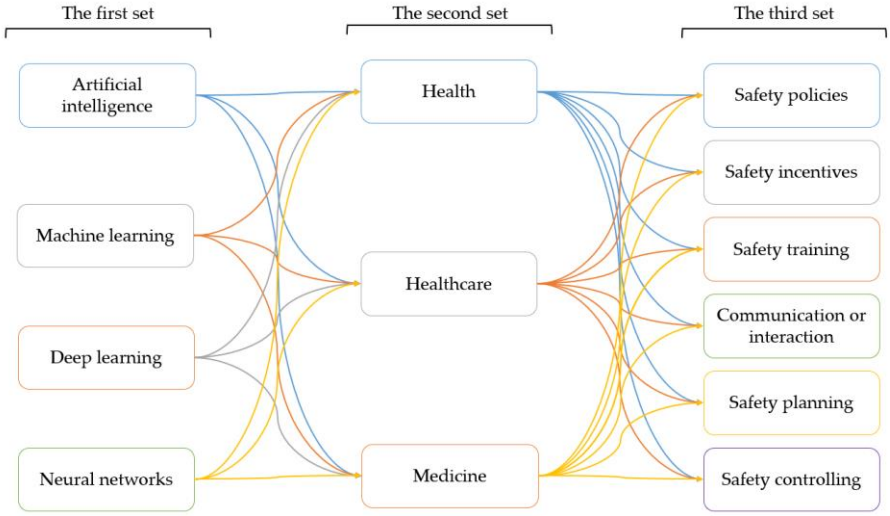

Figure 3. Three sets of keywords and their combinations for identifying relevant articles.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202012.0313.v2

Peer-reviewed version available at Symmetry 2021, 13, 102; doi:10.3390/sym13010102

The PubMed and Google Scholar databases were used to discover relevant articles published through the end of July 2020. The selection strategy is shown in Figure 4.

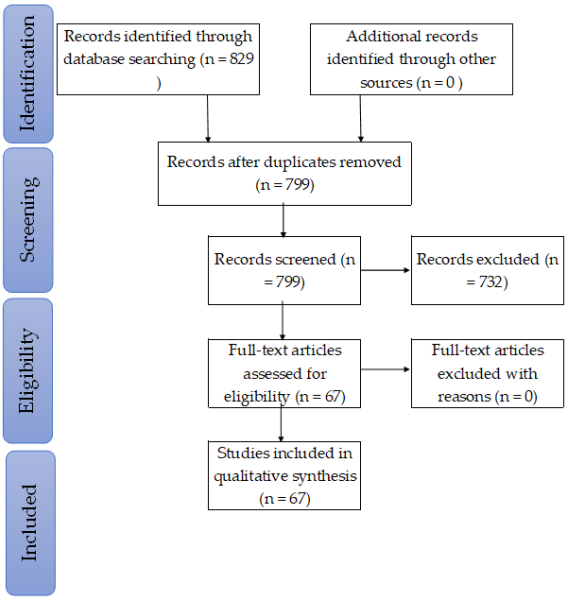

Figure 4. Chart of the selection strategy following PRISMA guidelines [38].

2.1.2. Interviews

A primary method for collecting qualitative data is interviews, which are widely used in conducting research [40]. Some researchers apply structured interviews to test a priori hypotheses by using standardized questions and analysis. In contrast, others apply qualitative interviewing methods to better understand the stated hypotheses [41]. In this article, both types of interviews were used.

After completing the systematic literature review, we applied an interview approach to identify the SCS's key dimensions. The objective was to examine extracted information from the literature review, discuss unidentified aspects of safety systems, and determine measurable third level attributes. Consequently, we interviewed ten AI domain experts. We asked these ten specialists the following questions:

- IRQ1. What are the attributes of safety policies for implemented AI models in healthcare?
- IRQ2. What are the attributes of incentives for clinicians for implemented AI models in healthcare?
- IRQ3. What are the attributes of clinician and patient training for implemented AI models in healthcare?
- IRQ4. What are the attributes of communication and interaction for implemented AI models in healthcare?
- IRQ5. What are the attributes of planning of actions for implemented AI models in healthcare?
- IRQ6. What are the attributes of control of actions for implemented AI models in healthcare?

The interviews were conducted during August 2020. Each interview lasted approximately 1 hour and was divided into two main parts. First, the key dimensions and results of the systematic literature review

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202012.0313.v2

Peer-reviewed version available at Symmetry 2021, 13, 102; doi:10.3390/sym13010102

were explained and discussed. Second, six research questions were asked, and the third level attributes were developed.

Because we conducted information-gathering interviews where IRQ1-6 questions only concentrated on AI-based models rather than individuals or their opinion regarding themselves; therefore, our study is not considered human subjects research, and ethics approval was not required [41]. However, before participating in the interview, we transparently informed all individuals about our study's objective and aim. We emphasized that participation is voluntary. Therefore, participants were free to leave the interview without any penalty or question. We ensured that individuals were not pregnant, consumed alcohol for 24 hours before the interview, or were under hormonal treatment. We sent a recruiting email accompanied by a list of questions (IRQ1-6) and an explanation of our research to all potential candidates. We did not collect any identifiable data. All emails, contact information, messages were deleted right after interviews. Finally, we obtained informed consent verbally from all individuals. The socio-demographic information of interviewees is shown in Table 1.

Table 1. The socio-demographics information of interviewees

Characteristics	Interviewees (number)	Interviewees (percent)
Age		
30 to 34	2	20%
35 to 39	4	40%
40 to 44	4	40%
Years of experience in AI		
0 to 4	1	10%
5 to 9	4	40%
10 to 14	5	50%
Gender		
Male	10	100%
Female	0	0
Race/Ethnicity category		
Non-Hispanic Black	0	0
Non-Hispanic Asian	0	0
Non-Hispanic White	10	100%
Non-Hispanic Other	0	0
Hispanic	0	0
Occupation		
Postdoctoral researcher	2	20%
Data scientist	5	50%
Machine learning scientist	2	20%
Data engineer	1	10%

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202012.0313.v2

Peer-reviewed version available at Symmetry 2021, 13, 102; doi:10.3390/sym13010102

After discovery of attributes on the basis of the systematic literature review and expert interviews, we organized the key dimensions and the second and the third level attributes into a hierarchy tree. In this knowledge structure, the higher-level attributes represented the overall view of safety in implemented AI models, and the lower-level attributes measured the elements of safety in AI models (Figure 5). Notably, the highest level had 6 attributes, the middle level had 14 attributes, and the lower level had 78 attributes.

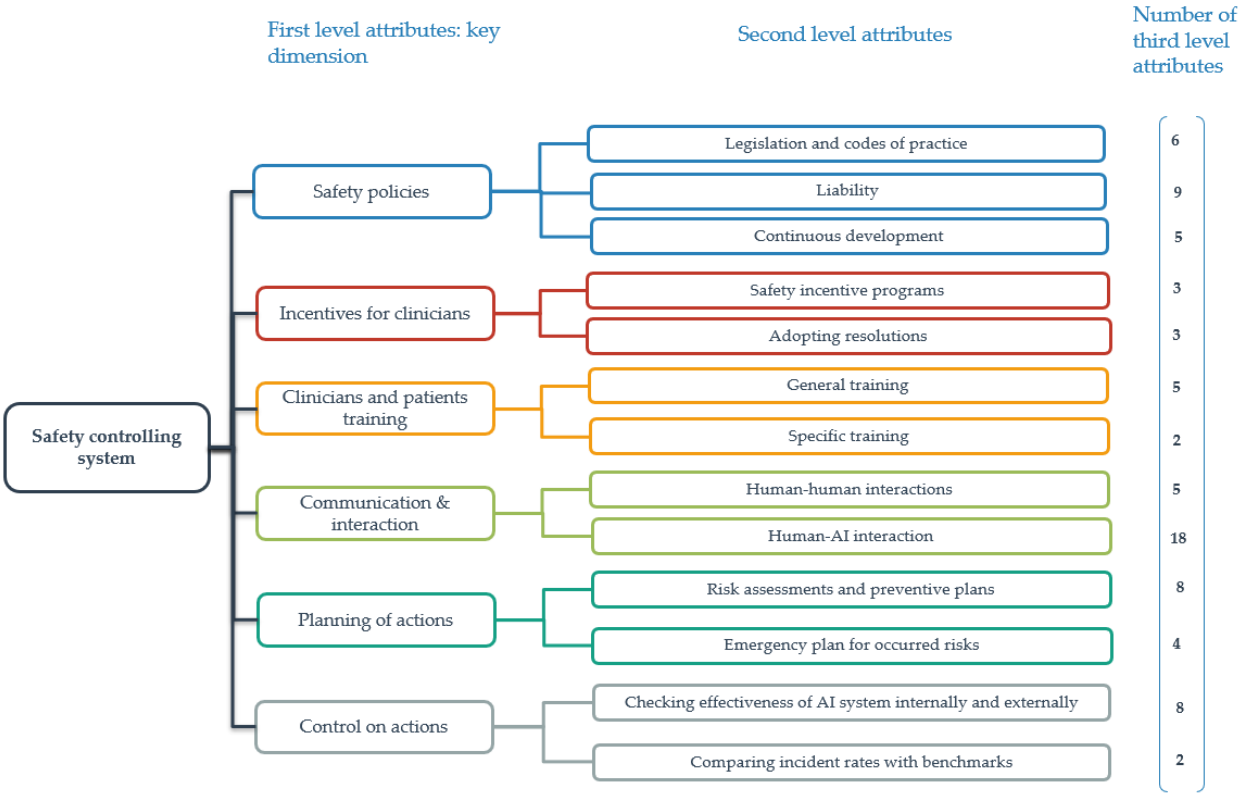

Figure 5. Attributes of the SCS in different layers.

2.2. Weight of attributes

Because the identified attributes differed in importance regarding AI system safety, differentiating essential attributes from desirable attributes was essential. Therefore, we assigned a weight to each attribute to understand its degree of importance. Weights are crucial for decision-making because they indicate the most critical safety elements in AI systems implementation. For assigning weights to attributes, we used a 4-point Likert scale. For this purpose, a questionnaire was designed containing the third level attributes. To evaluate the significance of the third level attributes, we asked ten AI experts who participated in developing the attributes to rate these attributes on a 4-point scale: not important = 1; neutral = 2; important = 3; and very important = 4.

We assessed agreement among AI experts by calculating Kendall's W (Kendall's coefficient of concordance) [43,44]. This non-parametric statistic ranges in value between 0 and 1, such that 1 indicates more substantial agreement [45]. We assessed the concordance of opinions regarding six key dimensions of the SCS. There was strong agreement (Kendall's W scale bigger than 0.6) among AI experts in the key

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202012.0313.v2

Peer-reviewed version available at Symmetry 2021, 13, 102; doi:10.3390/sym13010102

dimensions of "planning of actions" and "control of actions." In addition, the AI experts were moderately in agreement (Kendall's W scale between 0.3 to 0.6) regarding the remaining key dimensions. However, we decided to adopt the average experts' ratings as each third level attribute's weight. Next, the weights of all third-level attributes were recalculated such that the sum of all weights was 100. For this to be achieved, we added up rates of the scale for all third-level attributes; then, we divided the rate of each attribute by the sum of all attributes. In the final step, the weights of the second and the first level attributes were determined. For this purpose, we added up the weight of all third-level attributes corresponding to the first and second level attributes. According to the results, the key dimension of "communication and interaction" had the highest weight and was followed by "control of actions" and "safety policies." The weights of the key dimensions and the second level attributes are shown in Figure 6 and Figure 7.

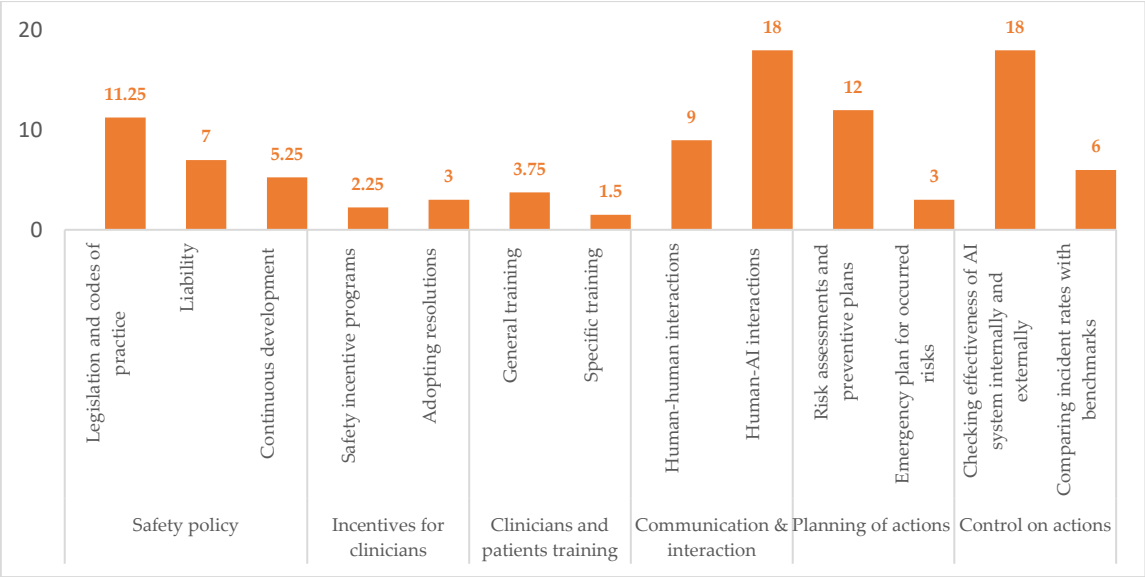

Figure 6. Weights of the first and second level attributes.

2.3. The rating system

The next part of the MAVT approach was developing a rating system and assigning it to the third level attributes. To reduce the probability of having different results from different auditors and to improve the generalization of the SCS, we developed a rating system by allocating points to the third level attributes in a straightforward manner. Different types of rating systems were extracted from Teo and Ling [36] and used in the survey. Four possible rating options were as follows:

1) 0/1, in which the rating options are "0" (no) or "1" (yes),

2) 0–1, in which the rating options are a fraction between "0" and "1",

3) 0/1/NA, in which the rating options are "0" or "1" or "not applicable", and

4) 0–1/NA, in which the rating options are a fraction between "0" and "1" or "not applicable."

A questionnaire was designed containing the third level attributes. To assign the most relevant rating system to each attribute, we asked the ten AI experts to select the most relevant rating system. We assessed the agreement among AI experts by assigning numbers from 1 to 4 to each rating system and calculating

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202012.0313.v2

Peer-reviewed version available at Symmetry 2021, 13, 102; doi:10.3390/sym13010102

Kendall's W. There was strong agreement (Kendall's W scale bigger than 0.6) among the AI experts in all key dimensions except "control of actions" and "safety policies." The most relevant (popular) rating system was assigned to each attribute according to the collected data.

2.4. Finalizing the model

The score of each third level attribute was determined by multiplying the attribute's weight in the auditor's assessment by the attribute according to the assigned rating system. After scores were determined, the total score was calculated by adding all scores of third level attributes. In conclusion, the entire MAVT approach for developing the SCS in 19 steps is represented in Figure 7.

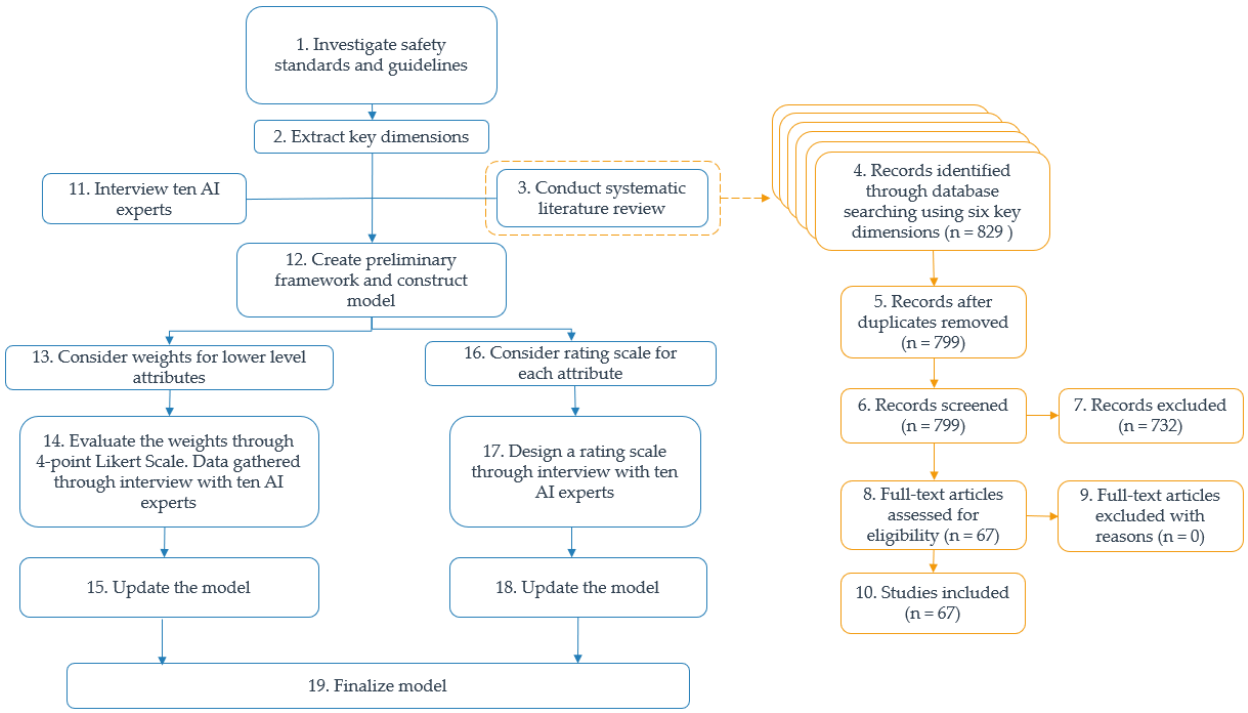

Figure 7. The 19 steps of the MAVT approach.

3. Results

An SCS can be defined as a set of policies, practices, procedures, strategies, roles, functions, and resources associated with safety that interacts in an organized way to decrease the damage generated in a process [37,46]. Different SCSs have been developed for different industries and technologies, but there is a lack of studies aiming to understand the key dimensions and measurable indicators of the safety of black-box AI models in healthcare. Although the developed safety models and guidelines for industries may not apply directly to AI models in healthcare, their methods and frameworks can be adapted to create a comprehensive safety system suitable for black-box AI models.

This article developed a system to evaluate the safety performance of AI models implemented in healthcare. The proposed system was constructed by applying the three-level MAVT approach [36]. The first level attributes, adopted from Fernández-Muñiz et al. [37], were the main elements of safety standards and guidelines. The 14 attributes of the second level and the 78 elements of the third level were extracted

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202012.0313.v2

Peer-reviewed version available at Symmetry 2021, 13, 102; doi:10.3390/sym13010102

by using a systematic literature review, conducting interviews, and performing two small questionnaire-based surveys.

The key first level dimensions of the SCS are as follows: (1) safety policies; (2) incentives for clinicians; (3) clinician and patient training; (4) communication and interaction; (5) planning of actions, and (6) control of actions (Figure 8).

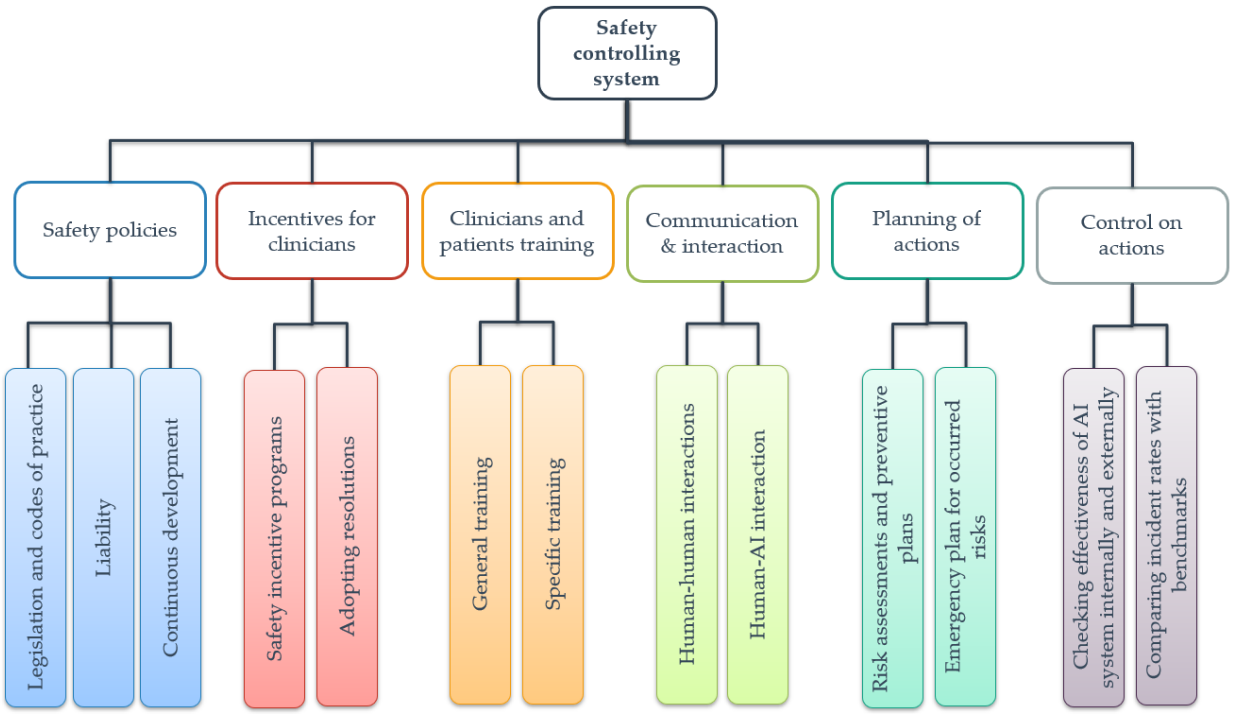

Figure 8. The first and second level attributes of the SCS.

The first key dimension.

Safety policies can be divided into three attributes of legislation and codes of practice (including six attributes), liability (including nine attributes), and continual development (including five attributes). Developing AI models in healthcare faces several legal regimes, such as federal regulations, state tort law, the Common Rule, and the Federal Trade Commission Act. In malpractice claims, owing to the use of AI black-box models in clinical workflows, the current legal system is not suitable [47,48]. Therefore, the responsibilities of different parties, including AI developers, the source of training data, clinicians, and suppliers who provide the AI system platform, must be clearly defined [1]. In addition, clinical systems are being controlled by designed rules, and using black-box data-driven devices can introduce new risks [49]. For example, traditional medical devices are updated manually, whereas AI-based devices are updated by training with new data [49]. The differences between the data-driven and the traditional devices require clinical regulations to be updated to correctly implement AI-based devices in the clinical workflows [47].

The continual development of AI models is one of the main attributes of safety policies. AI-based devices are a new type of medical technology, and they may become outdated because of continually changing medical treatment patterns and improvements in medical instruments [50,51]. Therefore, this unique aspect of AI-based medical devices in medical regulation is important to consider [1]. The Food and Drug Administration (FDA) has defined and developed the Software as Medical Device (SaMD) category

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202012.0313.v2

Peer-reviewed version available at Symmetry 2021, 13, 102; doi:10.3390/sym13010102

and Digital Health Software Precertification (Pre-Cert) Program to address this issue [47,52]. Accordingly, the FDA's new policy centers on AI developers' organizational excellence rather than approving AI-based medical devices [53]. Organizational excellence is defined as an effort to develop different processes and standards intended to engage employees to deliver excellent products [53]. Consequently, developers are authorized to update AI models without to review by the FDA [54]. However, a testing process must crucially be developed so that the updated models' performance is not below that of the primary models [55].

The second key dimension.

Because of false confidence, clinicians often accept the results and recommendations of AI models regardless of their accuracy [29]. To address this issue and to motivate clinicians to constantly check the results of AI systems, clinicians' incentives can be considered according to two attributes: developing safety incentive programs and adopting resolutions according to clinicians' recommendations. Safety incentive programs comprise three aspects, such as: "Are there any incentives offered to clinicians to put defined procedures of implemented AI systems into practice?" The attribute of adopting resolutions comprises three aspects, such as: "Is there any meeting with clinicians to adopt their recommendations concerning AI-based medical device operation?"

The third key dimension.

Clinicians and patients should be educated on the risks, benefits, and limitations of AI models [33]. Different actions that should be taken may include engaging clinicians in developing data-driven systems, providing training events in health organizations before and after AI model implementation, using different teaching methods to educate clinicians, asking for feedback from learners, and developing personalized education [56]. Clinician and patient training can be divided into (1) general training, including five attributes, such as "Are clinicians given sufficient training concerning AI system operation when they enter a health institution, change their positions or use new AI-based devices?" and (2) specific training for certain patients and clinicians facing high-risk events, including two attributes, such as "Are specific patients or clinicians trained who are facing high-risk events?"

The fourth key dimension.

Communication and interaction can be divided into two main attributes of human-human interactions: interaction between parties (such as healthcare institutions and AI developers, including five attributes), and human-AI interaction (including 18 attributes mainly adopted from Amershi et al.) [57]. The safe implementation of AI-based devices in healthcare depends on comprehensive and effective interaction among healthcare institutions, clinicians, and AI developers. This interaction is necessary, because AI models cannot be trained and tested for all disease states and patient demographics during clinical trials. This interaction includes five attributes, such as "Is there any information system developed between a health institution and an AI developer during the lifetime of AI-based medical devices?" [58–60]. The attribute of human-AI interaction is associated with designing safety guidelines for interaction between humans and AI. A total of 18 attributes are considered for human-AI interaction, such as: "Is there any established description of what the AI-based medical device can do?" [57].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202012.0313.v2

Peer-reviewed version available at Symmetry 2021, 13, 102; doi:10.3390/sym13010102

The fifth key dimension.

Because implemented AI models are required to perform tasks in dynamic and complex healthcare environments, and because these models cannot be fully evaluated during clinical trials, a safety plan must be created to identify all risks and adverse events, and plan a course of action to remove risks and plan emergency actions. The planning of actions is divided into two main aspects: (1) risk assessment and preventive plans, including eight attributes, such as: "Are all risks and adverse events identified concerning implemented AI systems?" and (2) emergency plan for occurred risks, including four attributes, such as: "Do the health institution and the AI developer have an emergency plan for remaining risks and adverse events of AI operation?"

The sixth key dimension.

The control of actions aims to monitor all risks and adverse events and all procedures and planning. The control of actions is divided into two main aspects: 1) checking the AI system's effectiveness internally and externally, and 2) comparing incident rates with benchmarks. Checking the AI system's effectiveness internally and externally involves eight attributes, such as "Is effective post-market surveillance developed to monitor AI-based medical devices?" Post-market surveillance has two main parts. The first is practical cooperation among clinicians, health organizations, and AI developers to gather clinical and safety-related data (explained the communication and interaction attribute). The second is monitoring and analyzing different safety signals, longitudinal data, risks and adverse events, and thresholds for AI-based device recall [61]. The comparison of incidence rates with benchmarks includes two attributes. All attributes are shown in Table 2.

Table 2. Summary of system attributes.

Attributes	Weight	Rating system
SCS	100.00	
Safety policies	23.50	
Legislation and codes of practice	11.25	
Is there a commitment to current legal regimes, such as federal regulations, state tort law, the Common Rule, Federal Trade Commission Act, legislation associated with data privacy, and legislation associated with the explainability of AI?	3.00	0–1
Is a written declaration available reflecting the safety objectives of the AI-based medical device?	2.00	0/1
Are clinicians informed about the safety objectives of the AI-based medical device?	2.00	0/1
Is a written declaration available reflecting the safety concerns of the directors of health institution?	1.50	0/1
Does the health institution coordinate the AI-based medical device policies with other existence policies?	1.50	0/1
Is there a positive atmosphere to ensure that individuals from all parties, such as the health institution and the AI developer, participate in and contribute to safety objectives?	1.25	0–1
Liability	7.00	
Are the responsibilities of the AI developer established in writing?	1.00	0/1
Are the responsibilities of clinicians established in writing?	0.75	0/1
Are the responsibilities of the source of training data established in writing?	0.75	0/1
Are the responsibilities of the source of suppliers who provide the system platform established in writing?	0.75	0/1
Are the responsibilities of the AI algorithm (at the higher level) established in writing?	0.25	0/1/NA

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202012.0313.v2

Peer-reviewed version available at Symmetry 2021, 13, 102; doi:10.3390/sym13010102

Attributes	Weight	Rating system
Is there a positive atmosphere to ensure that individuals from all parties, such as the health institution and the AI developer, know their responsibilities?	0.75	0–1
Is there an appropriate balance for the responsibilities of different parties?	0.75	0–1
Is there any procedure for resolving conflicts between parties?	1.00	0/1
Are resolving conflicts established in writing?	1.00	0/1
Continuous development	5.25	
Is there a commitment to FDA regulations regarding Software as Medical Device (SaMD)?	0.75	0/1
Is there involvement in the Digital Health Software Precertification (Pre-Cert) Program?	0.75	0/1/NA
Is an organizational excellence framework established in writing?	0.75	0/1
Is there a commitment to organizational excellence?	1.50	0/1
Is there a testing policy for updated AI-based devices?	1.50	0/1
Incentives for clinicians	5.25	
Safety incentive programs	2.25	
Are there any incentives offered to clinicians to put defined procedures of implemented AI systems into practice?	0.75	0/1
Are incentives frequently offered to clinicians to suggest improvements in the performance and safety of implemented AI systems?	1.00	0/1
Are there disincentive programs for clinicians who fail to put defined procedures of implemented AI systems into practice?	0.50	0/1
Adopting resolutions	3.00	
Are there any meetings with clinicians to adopt their recommendations concerning AI-based medical device operation?	1.50	0/1
Is adoption of resolutions coordinated with other parties, such as the AI developer?	0.50	0/1
Do any modifications or changes in AI-based medical device operations involve direct consultation with clinicians who are affected?	1.00	0/1
Clinician and patient training	5.25	
General training	3.75	
Are clinicians given sufficient training concerning AI system operation when they enter a health institution, change their positions or use new AI-based devices?	1.75	0/1
Is there a need for follow-up training?	0.50	0/1/NA
Are general training actions continual and integrated with the established training plan?	0.50	0/1/NA
Are the health institution's characteristics considered in developing training plans?	0.50	0/1/NA
Is the training plan coordinated with all parties, such as the AI developer and health institution?	0.50	0–1/NA
Specific training	1.50	
Are specific patients or clinicians trained who are facing high-risk events?	0.75	0/1/NA
Are specific training actions continual and integrated with the established specific training plan?	0.75	0/1/NA
Communication and interaction	27.00	
Human-human interactions	9.00	
Is any information system developed between a health institution and an AI developer during the lifetime of AI-based medical devices?	2.00	0/1
Are clinicians informed before modifications and changes in AI-based medical device operation?	2.00	0/1
Is there written information about procedures and the correct way of interacting with AI-based medical devices?	2.00	0/1
Is there any communication plan established between parties?	1.50	0–1
Is there any procedure to monitor communication and resolve problems such as language, technical, and cultural barriers between parties?	1.50	0/1
Human-AI interactions	18.00	
Is there any established description of what the AI-based medical device can do?	1.50	0/1
Is there any established description of how well the AI-based medical device performs?	1.50	0/1
Is the AI-based medical device time service (when to act or interrupt) based on the clinician's current task?	1.50	0/1

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202012.0313.v2

Peer-reviewed version available at Symmetry 2021, 13, 102; doi:10.3390/sym13010102

Attributes	Weight	Rating system
Does the AI-based medical device display information relevant to the clinician's current task?	1.50	0/1
Are the clinicians interacting with AI-based medical devices in a way that they would expect (are social and cultural norms considered)?	1.50	0/1
Is there any procedure to ensure that the AI-based medical device's behaviors and language do not reinforce unfair and undesirable biases?	1.50	0/1
Is it easy to request the AI-based medical device's services when needed?	0.75	0/1
Is it easy to ignore or dismiss undesired and unwanted AI-based medical device services?	0.75	0/1
Is it easy to refine, edit, or even recover when the AI-based medical device is wrong?	0.75	0/1
Is it possible to disambiguate the AI-based medical device's services when they do not match clinicians' goals?	0.75	0/1
Is it clear why the AI-based medical device did what it did (access to explanations and visualizations of why the AI-based medical device behaved as it did, in terms of mitigating the black-box)?	0.75	0/1
Does the AI-based medical device have short term memory and allow clinicians to efficiently access the memory?	0.75	0/1
Does the AI-based medical device learn from clinicians' actions (personalizing clinicians' experience by learning from their behaviors over time)?	0.75	0/1
Are there several disruptive changes when updating the AI-based medical device?	0.75	0/1
Can clinicians provide feedback concerning the interaction with the AI-based medical device?	0.75	0/1
Can the AI-based medical device identify clinicians' wrong or unwanted actions? How it will react to them?	0.75	0/1
Can the clinicians customize what the AI-based medical device can monitor or analyze?	0.75	0/1
Can the AI-based medical device notify clinicians about updates and changes?	0.75	0/1
Planning of actions	15.00	
Risk assessments and preventive plans	12.00	
Are all risks and adverse events identified concerning the implemented AI system?	2.50	0/1
Is there any system in place for assessing all detected risks and adverse events of AI operation?	1.75	0/1
Are prevention plans established according to information provided by risk assessment?	1.75	0/1
Does the prevention plan clearly specify for clinicians who are responsible for performing actions?	1.25	0/1
Are specific dates set for performing preventive measures?	1.25	0/1
Are procedures, actions, and processes elaborated upon on the basis of performed preventive measures?	1.50	0/1
Are clinicians (involved in using the implemented AI system) informed about prevention plans?	1.00	0/1
Are prevention plans occasionally reviewed and updated on the basis of any changes or modifications in operation?	1.00	0/1
Emergency plan for risks	3.00	
Is an emergency plan in place for the remaining risks and adverse events of AI operation?	0.75	0/1
Does the emergency plan clearly specify for clinicians who are responsible for performing actions?	0.75	0/1
Are the clinicians (involved in using the implemented AI system) informed about the emergency plan?	0.75	0/1
Is the emergency plan occasionally reviewed and updated on the basis of any changes or modifications in operation?	0.75	0/1
Control of actions	24.00	
Checking the effectiveness of AI system internally and externally	18.00	
Is effective post-market surveillance developed to monitor AI-based medical devices?	2.50	0/1/NA
Are there occasional checks performed on the execution of the preventive plan and emergency plan?	2.50	0/1
Are there procedures to check collection, transformation, and analysis of data?	2.25	0/1

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202012.0313.v2

Peer-reviewed version available at Symmetry 2021, 13, 102; doi:10.3390/sym13010102

Attributes			Weight	Rating system
		Is there a clear distinction between the information system and the post-market surveillance system?	2.25	0/1
		Are accidents and incidents reported, investigated, analyzed, and recorded?	2.25	0/1
		Are there occasional external evaluations (audits) to validate preventive and emergency plans?	2.00	0/1/NA
		Are there occasional external evaluations (audits) to ensure the efficiency of all policies and procedures?	2.00	0/1/NA
		Are there procedures to report the results of external and internal evaluation?	2.25	0/1/NA
	Comparing incident rates with benchmarks		6.00	
		Do the accident and incident rates regularly compare with those of other healthcare institutions from the same sector using similar processes?	3.00	0/1/NA
		Do all policies and procedures regularly compare with those of other healthcare institutions from the same sector using similar processes?	3.00	0/1/NA

4. Discussion

This study offers an alternative solution for opening the AI black-box in healthcare by introducing an SCS. The framework provides safety guidelines for implementing AI black-box models to reduce the risk of healthcare-related incidents and accidents. The proposed framework and system provide a basis for implementing and monitoring safety legislation and procedures, identifying the risks and adverse events in AI activities, preventing accidents and incidents from occurring, and having an emergency plan for threats. Therefore, the proposed framework and tool can guide the safety activities of implemented AI systems.

The SCS represents a set of attributes in different layers and can be used in healthcare institutions with implemented AI models. The management of healthcare institutions can use the proposed set of attributes as a checklist, verifying whether a set of desired safety elements exists. Having useful specific attributes in healthcare systems will lead to high scores in the SCS. Healthcare institutions can use this framework to (1) calculate their safety score, and compare it with those of other institutions, and (2) detect deficiencies in current safety practices regarding the implemented AI models. The above steps can help improve the overall safety performance.

The proposed framework for evaluating AI safety performance was developed by using the MAVT approach, comprising four parts: extracting attributes, generating weights for attributes, developing a rating scale, and finalizing the system. With the MAVT approach, three layers of attributes were created. The first level contained 6 key dimensions, the second level contained 14 attributes, and the third level contained 78 attributes.

First key dimension.

Three attributes—"legislations and codes of practice," "liability," and "continual development"—were extracted as primary elements of safety policies from the literature review and were confirmed in interviews. Commitment to current legislation and codes of practice is a basic element of every AI system. Among current legal regimes, data privacy-related legislation plays a vital role in developing and implementing AI systems. Because of the complexity of protecting data privacy and its effects on data

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202012.0313.v2

Peer-reviewed version available at Symmetry 2021, 13, 102; doi:10.3390/sym13010102

availability, three different viewpoints concerning the level of adaptation of data protection legislation have recently been proposed.

First view. The European Union has adopted legislation entitled General Data Protection Regulation (GDPR), which details a comprehensive and uniform approach for data privacy, regardless of how data are collected, in what format, or who the custodian is [62]. Under GDPR, only anonymous data can be shared. The anonymization process under GDPR requires implementing different techniques on datasets to prevent data re-identification [62]. Although GDPR aims to protect data privacy rather than to prevent data sharing, a fear of violation penalties has decreased data collection and data aggregation efforts among European companies and even data flow from Europe to the U.S. [63].

Second view. The current U.S. data privacy legislation is more lenient than that of the European Union [64]. In general, Europe places more emphasis on protecting citizens from technological risks, whereas the U.S. focuses more on innovation and technology [64]. Under U.S. privacy law, health data are treated differently depending on how they have been created, who is handling the data, and who the data custodian is [65]. The Health Insurance Portability and Accountability Act (HIPAA) includes a privacy rule that prohibits disclosing protected health information [47]. HIPAA limits the use of protected health information unless there is authorization from the patient or Institutional Review Board [65]. Under HIPAA policy, any type of de-identified data is considered non-personal and not subject to data protection regulation [62]. Furthermore, HIPAA focuses on specific actors and their activities rather than on the data itself; therefore, a considerable amount of health data are not covered by HIPAA [65].

Third view. From China's perspective, AI is a powerful tool for economic success, military dominance, and controlling the population [63]. Chinese companies accumulate a tremendous amount of health-related data, which can be used in AI development, owing to lenient regulation on data collection and little public concern about data privacy [54,66]. However, in recent years, the Chinese public has started to petition large companies, such as Baidu and Alibaba, for the right to data privacy [66]. Consequently, China has initiated personal data protection laws and ethical principles for developing and using AI [67,68].

Among the third level attributes of safety policies, the elements "Software as Medical Device (SaMD)," "Digital Health Software Precertification (Pre-Cert) Program," "current legal regimes," and "assigning responsibility" were mainly extracted from the included articles. The elements of "safety objectives of the AI-based medical device," "positive atmosphere in the health institution," and "coordinating the AI-based medical device policies with existing policies" were mainly found from the interviews. However, we observed that the AI experts differed in the weights assigned to this crucial dimension's attributes. The most confusing second level attribute was the liability, on which AI experts did not reach agreement.

The term *Software as a Medical Device (SMD)* is described as "software that uses an algorithm that operates on data input to generate an output that is used for medical purposes" [69]. SMD applications are diverse as Computer-Aided Detection (CAD) software—for example, software detecting breast cancer—and smartphone applications for diagnostic purposes—for example, software for analyzing images collected from a magnetic resonance imaging medical device. Although some FDA guidelines for SMD overlap with attributes of other key dimensions, we decided to consider "commitment to FDA regulations regarding Software as Medical Device" under "safety policies." As described earlier, the Pre-Cert Pilot Program looks first at the AI developers rather than at AI-based medical devices, in contrast to the FDA

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202012.0313.v2

Peer-reviewed version available at Symmetry 2021, 13, 102; doi:10.3390/sym13010102

process for traditional medical devices [70]. Because the FDA selected several companies to participate in developing the Software Pre-Cert pilot program, we decided to include it as an attribute.

Second and third key dimensions.

Both the "incentives for clinicians" and "clinician and patient training" attributes were formed and developed in interviews. There was moderate agreement regarding the weights of attributes and strong consensus regarding the assigned rating system.

Fourth key dimension.

Although two parts of this key dimension were mainly extracted from the literature review, a considerable amount of interview time was spent on this aspect to define the third level attributes. Human-human interactions are associated with communication management among all parties—for example, AI developers and health institutions—involved in implemented AI-based medical devices. All main communication management elements, including planning, managing, and monitoring communication, were discussed in interviews, and measurable attributes were defined. One of the main attributes of human-human interaction is developing an information system for storing, processing, collecting, creating, and distributing information. This information system contains different elements of hardware and software, system users and developers, and the data itself.

Regarding human-AI interactions, the attributes from Amershi et al. [57] were discussed in the interviews to define measurable attributes. The main elements of the human-AI interaction included the following: AI system capability, AI system accuracy, AI system time service, AI system displaying information, AI system language, social and cultural norms in human-AI interaction, AI system readiness, dismissal of unwanted service, AI system recovery, AI system disambiguation, AI system explainability (black-box mitigation), AI system short term memory, personalizing the AI system, updating the AI system, feedback mechanisms in the AI system, the AI system's reaction to wrong actions, customizing the AI system, and notification mechanisms in the AI system. Importantly, personalization means that AI systems can learn from clinicians' actions, and customization means that clinicians can customize the AI system's actions.

One of the main controversial elements of human-AI interaction is the AI system's accuracy and effectiveness. As a part of model safety, the AI model's performance in clinical trials should outperform the performance of existing diagnostics devices and clinicians' judgment [47]. Accuracy, defined as a proper fraction of predictions, is a commonly used metric for evaluating AI algorithms' performance [47].

Many studies have reported the three measures of accuracy, sensitivity, and specificity in clinical trials to capture the full extent of models' properties [47]. However, covering all essential differences in patient demographics and disease states in clinical trials is impossible [50]. One solution is to add external validation after the clinical trials before implementing the model in clinical workflows [50]. The external validation phase would include training and testing the model by using data from the clinics where the AI model will be used [50].

Other metrics to measure model performance are stability and robustness [35]. Model stability means that, when given two almost identical input data sets, an AI model generates almost the same results [71]. Model robustness indicates the stability of the model's performance after including noise in the input data

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202012.0313.v2

Peer-reviewed version available at Symmetry 2021, 13, 102; doi:10.3390/sym13010102

[35]. Robustness represents the model efficiency for new data outside the training data [35]. These measures are essential for applying AI models in healthcare, because the lack of stability and robustness can diminish clinicians' and patients' trust in AI models [72].

Fifth key dimension.

In this key dimension, risk assessment was mainly extracted from the literature review, and elements of the preventive plan and emergency plan were discussed in interviews. The foundation of the "planning of action" dimension is risk assessment. The principal risks of implemented AI systems include data difficulties, technological problems, security problems, models misbehaving, and interaction issues [73]. Two elements of models misbehaving and interaction issues were addressed in AI-human interactions. Therefore, the main risks associated with the implemented AI system are data difficulties and technology problems.

Risk of data difficulties. One of the main concerns regarding AI in healthcare is data availability [1,2,21]. Despite considerable recent efforts in collecting and releasing high-quality AI-ready datasets, most health data are not accessible to the public [1,2,21]. These data are generally collected and controlled by hospitals and other health organizations and used for operations but not for analytics or research. Therefore, the formats of the data are often not ideal for training AI models. For example, image data may not be anonymized, organized, or appropriately annotated [74]. Of the publicly available datasets, most are released once and become progressively outdated [50]. For example, despite advances in fundus camera technology, the Messidor database is still used to train AI algorithms on images acquired in 2007 [75].

Other issues in data availability include coverage of rare and novel cases [76], missing data in datasets, a lack of appropriately labeled data [77], high-dimensionality together with small sample sizes [78], and data contamination with artifacts and noise [79]. Among image datasets, the main issues include difficulty in collecting many high quality manually annotated images [80], the limitations of human perception in annotating and labeling images [81], the time required for reviewing and annotating each image in a dataset [82], the level of raters' sensitivity to a particular target [83,84],, loss of information due to image processing and resizing [85], and collection of images from only a specific device [86].

Data privacy is the main difficulty in increasing data availability in healthcare [1,2,21]. A delicate balance must be struck between stimulating the potential benefits of aggregating health data and protecting individual privacy rights. To do so, different reported practices include anonymizing data before sharing, using validated protocols for de-identification, exploring safer ways to share data, and defining the responsibilities of health organizations as data custodians [87]. However, linking de-identified data is much more difficult when patients visit different health institutions, obtain insurance through various companies, or change their location [65]. Consequently, forming fragmented health data makes data-driven innovation more difficult [65].

Mitigating the risk of data difficulties. High-quality AI-ready data are the foundation for developing accurate algorithms. Even the unintentional effects of biases due to selecting unsuitable data can decrease the accuracy of AI models. To generate high-quality AI-ready data, different methods have been proposed in various studies. Data aggregation efforts across health organizations are one way to generate high-quality data [88]. One of the main challenges in data aggregation is that the data format may differ among health organizations [89]. Therefore, usable data with consistently structured formats must be generated

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202012.0313.v2

Peer-reviewed version available at Symmetry 2021, 13, 102; doi:10.3390/sym13010102

among health organizations [89]. Several efforts have been proposed to address this concern, including developing cloud infrastructures, adopting unified data formats such as Faster Healthcare Interoperability Resources, and launching collaborative efforts among health organizations to create high-level joint features [1,90].

Training AI models in a simulated virtual environment has created a unique opportunity to cover the lack of high-quality healthcare data [91]. By using the virtual environment, an AI model can learn and become powerful before it is implemented in the physical world [92]. Chawla [93] has reported the successful implementation of AI models trained inside a virtual environment. The key advantages of using a virtual environment for training AI models are as follows:

1. The virtual environment allows AI developers to simulate rare cases for training models [92].
2. The entire training process can occur in a simulated environment without the need to collect data [93].
3. Learning in the virtual environment is fast; for example, AlphaZero, an AI-based computer program, was trained over a day to become a master in playing Go, chess, and shogi [29].

However, using a virtual environment for training AI models in healthcare is not as advanced as its applications in other fields, such as autonomous cars. For example, the Waymo company has created virtual models of whole cities, and every day it sends 25,000 virtual self-driving cars through these cities to train AI algorithms [94]. Using a virtual environment gives Waymo the ability to simulate more than 5 billion miles of autonomous driving [94]. This achievement may inspire healthcare companies to develop a vast virtual world including all disease states, patient demographics, and health conditions to train AI models.

Another way of generating high-quality data is building health datasets comprising data from volunteers and groups of consenting individuals. Encouraging patients to share their electronic medical record information and medical images, and creating datasets of volunteers' data have been described in several studies [74]. For example, in 2015, the U.S. National Institutes of Health set an objective to develop genomic data, lifestyle data, and biomarker data from 1 million volunteers from diverse backgrounds [54]. Another project supported by Google is developing a dataset comprising data from 10,000 volunteers over 4 years [89]. Participants in this project monitor their sleeping patterns and daily activities, answer common questions, and periodically visit specific medical testing locations [89]. However, various concerns exist regarding this type of data generation, including the lack of a specific mechanism for patients to share their data and the absence of a well-founded repository for aggregating patient data outside health care organizations. Awareness about the benefits of this process is lacking, and no institution has been authorized to monitor these projects [74,95].

The involvement of tech companies in healthcare has created a new trend of high-quality data generation [96]. For example, big tech companies collect massive amounts of behavioral data from social media and sensors [96]. Biomedical signals such as heart rate and rhythm, blood pressure, blood oxygen saturation, voice, tremor respiratory rate, limb movement, and temperature can be recorded by modern wearable devices [21]. These biological signals can be used for detecting several health conditions and diseases [2]. Patient-generated health data are another unique method for creating high-quality data. Various health-related datasets can be built by patients and caregivers outside clinics by using software applications, wearable sensors, monitoring devices, smartphones, and tablets with cameras [97]. Recently,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202012.0313.v2

Peer-reviewed version available at Symmetry 2021, 13, 102; doi:10.3390/sym13010102

substantial improvements have been made in high-quality and low-cost technologies with the potential to collect various patient-generated data regarding movement and behavior, environmental toxins, social interactions, diseases, images, and other physiological variables [98]. For example, one study has begun developing comprehensive open-access datasets through parents recording the behavior of their children with autism by using cell phone cameras [97]. In addition, the FDA has made efforts to establish a path for collecting patient- and caregiver-generated health data in clinical trials [99].

Collecting lifespan data from implemented AI-based medical devices is another method to access high-quality health data. These efforts require creating a system as a combination of hardware and software components to store and transfer generated data [100]. For example, by implementation of an AI model in different health organizations, high-quality data can be collected and stored in a repository outside health organizations, with consideration of data privacy protection [1].

Security problems. One of the main risks associated with implemented AI systems is security. Adversarial attacks, one of the major types of security problems in the AI system, can result when flawed AI systems are susceptible to manipulation by inputs explicitly designed to fool it [50]. For example, one study has shown that adding a very small amount of perturbation to images can cause medical image classifiers to incorrectly classify a mole with a 100% confidence level [50]. Because the issue of adversarial attacks cannot be completely addressed in clinical trials, fully managing malicious attacks is a main aspect of the safe implementation of AI systems in healthcare. Hostile attacks can be partially addressed by effective post-market surveillance; however, implementing regulatory actions and novel techniques can secure AI systems against adversarial attacks [11]. For example, in situations in which clinical data can be changed with fraudulent intent, using the BlockChain technique allows for data storage in immutable interconnected blocks [11].

Technological problems. Typically, the technological problems in AI systems relate to software and hardware. From a software perspective, AI systems are explicitly concerned with algorithms. Although we have discussed the main issues associated with algorithms, such as data difficulties and accuracy, generalization and algorithm fairness must also be addressed. Unknown accuracy of the results for minority subgroups is a major element of algorithm fairness [50]. For example, one study has developed an AI algorithm with high accuracy in the classification of benign and malignant moles but has found that it has poor performance on images of darker skin because it was trained on data from mainly fair-skinned patients [50]. Therefore, in developing and implementing AI systems, further training of AI models on data from minority groups, and the accuracy of AI models for underrepresented groups, must be considered [50].

From a hardware perspective, AI systems are mostly concerned with implementing algorithms on a physical computation platform [101]. Different physical computation platforms—distinguishable in terms of power efficiency, computation capability, and form factor—have been developed for AI systems, including a general-purpose central processing unit; graphical processing units; customizable and programmable accelerator hardware platforms, such as application-specific integrated circuits and field-programmable gate arrays; and other emerging platforms such as memristor crossbar circuits [101]. However, from the hardware perspective, the memory wall is a major challenge for AI systems [101]. The memory wall is defined as a situation in which improvements in processor speed are masked by the much

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202012.0313.v2

Peer-reviewed version available at Symmetry 2021, 13, 102; doi:10.3390/sym13010102

slower progress in dynamic random access (DRAM) memory speed [102]. Although DRAM organization has improved, this aspect is a major issue in AI systems [102].

Sixth key dimension.

Among elements of this key dimension, post-market surveillance was mainly extracted from the literature review. This effort was supplemented by internal and external validation and use of benchmarks formed and discussed during interviews. Part of the safe implementation of AI-based healthcare devices is post-market surveillance to monitor medical devices' safety [61]. Implementing comprehensive and effective post-market surveillance is essential for two reasons: (1) the FDA's new policy focuses on AI developers rather than AI-based medical devices, and (2) AI models cannot be trained and tested for all disease states and patient demographics during clinical trials and external validation [59]. The post-market surveillance system should include practical cooperation among clinicians, health organizations, and AI developers to efficiently gather clinical and safety-related data. Such a system should correctly identify safety signals, practically collect longitudinal data, effectively report adverse events, and strictly define thresholds for device recall [61,103]. An ideal level of post-market surveillance in AI-based medical devices includes three parts: extensively collecting data across the lifespan of devices, integrating results into electronic health records, and full tracking and reporting adverse events [58]. Developing and implementing a clear definition and distinction between information systems (data for human-human interaction) and post-market surveillance systems (data for AI-human interaction) is crucial.

5. Study limitations

The proposed framework of the AI SCS in the healthcare industry has several limitations. First, we did not perform safety audits to ensure the developed tool's effectiveness. Therefore, at this time, the quality of the proposed approach cannot be assessed in terms of:

1. The comprehensibility of the considered safety elements to potential auditors.
2. The robustness of the rating scale for each safety element to secure a reliable rating under similar conditions.
3. The potential for improving key dimensions and different layers of attributes.
4. The feedback from the healthcare institutions about the system.

To address the above challenges, the proposed framework should be implemented in several healthcare institutions concurrently to investigate its effectiveness. In addition, several key questions should be addressed, including (1) clinicians' acceptance of the framework, (2) the compatibility of the model across multiple healthcare institutions, (3) the opportunity for implementation in different types of healthcare organizations, (4) and the framework's effectiveness.

The second limitation of this study is the number of interviewees and their socio-demographic information. Many attributes were identified during the interviews, thus indicating their importance in developing AI safety system requirements. However, we interviewed ten AI experts who were middle-aged white males. Therefore, the small number of interviewees and their lack of diversity can introduce potential bias into the developed attributes.

Finally, the structural relationships between measurable variables (the third level attributes) and latent variables (the first and the second level attributes) should also be assessed to validate the developed model.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202012.0313.v2

Peer-reviewed version available at Symmetry 2021, 13, 102; doi:10.3390/sym13010102

A survey including many health institutions considering implementation of AI-based systems (including medical devices) should be conducted for that purpose. Another essential consideration is developing a set of robust AI-relevant safety criteria. Finally, implementation of the proposed system in real settings would require comprehensive management and appropriate regulatory oversight.

6. Conclusion

This article has discussed the challenges in advancing the implementation of AI in healthcare. We have outlined the safety challenges of AI in the context of explainability as opposed to the black-box approach. Our main objective was to propose a framework for controlling AI systems' safety as an alternative to opening the black-box. We adopted the MAVT approach to develop an AI system's safety attributes at three levels. This development process consisted of four parts: extracting attributes, generating weights for attributes, creating a rating scale, and finalizing the framework's architecture. We used a systematic literature review and interviews with subject experts to establish the safety attributes' hierarchical structure. We integrated the systematic review and interviews to understand better the main aspects of AI safety in the published literature and extend these aspects through consultation with AI domain experts. The first level contained 6 key dimensions, the second level included 14 attributes, and the third level had 78 attributes. Questionnaire-based surveys were used for assigning the weights and developing the attribute rating system. Finally, the limitations of the proposed AI safety controlling framework were discussed.

The first level key dimensions of the SCS are as follows: (1) safety policies; (2) incentives for clinicians; (3) clinician and patient training; (4) communication and interaction; (5) planning of actions; and (6) control of actions. In safety policies, it is essential to pay extra attention to the adaptation of data protection legislation. Because of the complexity of data privacy, many countries adapted their data protection legislation. In safety policies, the elements "Software as Medical Device (SaMD)," "Digital Health Software Precertification (Pre-Cert) Program," were discussed in detail by included articles.

The key dimension of communication and interaction can be divided into two main elements of human-human interactions and human-AI interactions. For the human-human interaction, it is necessary to develop an information system for storing, processing, collecting, creating, and distributing information. Several elements must be addressed for the human-AI interactions, such as AI system capability, AI system accuracy, and AI system explainability (black-box mitigation). Among elements of human-AI interaction, included papers discussed the AI system's accuracy and effectiveness.

In the key dimension of planning of actions, the principal risks of AI systems include data difficulties, technological problems, security problems, and models misbehaving. In data difficulties, data privacy is the main problem for increasing data availability in healthcare. However, new approaches are being developed to increase data availability in the healthcare sector, including data aggregation efforts across health organizations, training AI models in a simulated virtual environment, building health datasets comprising data from volunteers and groups of consenting individuals, the involvement of tech companies in healthcare, collecting lifespan data from implemented AI-based medical devices, and patient-generated health data. The adversarial attacks are one of the major security problems of the AI system. The technological problems in AI systems can be divided into software and hardware.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202012.0313.v2

Peer-reviewed version available at Symmetry 2021, 13, 102; doi:10.3390/sym13010102

Concerning the control of actions, it is necessary to have effective post-market surveillance to monitor medical devices' safety. As a part of this system, it is necessary to have practical cooperation among clinicians, health organizations, and AI developers to gather clinical data.

The implementation of the proposed framework in healthcare institutions should allow understanding its effectiveness better. In the near future, the key questions concerning this framework should also be addressed, including (1) clinicians' acceptance of the framework, (2) the compatibility of the model across multiple healthcare institutions, (3) the opportunity for implementation in different types of healthcare organizations. Also, we encourage other researchers to assess the structural relationships between measurable variables (the third level attributes) and latent variables (the first and the second level attributes) to validate the developed model.

Author Contributions: M.D.: conceptualization, methodology, and writing—original draft and revisions. W.K.: conceptualization, methodology, and writing—original draft, editing, and final revisions. K.F.: conceptualization, methodology, editing and revisions. T.W..: conceptualization, editing and final revisions. H.P.: conceptualization, editing and final revisions.

Funding: This research received no external funding.

Conflicts of interest: The author declares no conflict of interest.

References:

1. He, J.; Baxter, S.L.; Xu, J.; Xu, J.; Zhou, X.; Zhang, K. The Practical Implementation of Artificial Intelligence Technologies in Medicine. *Nat. Med.* **2019**, *25*, 30–36, doi:10.1038/s41591-018-0307-0.
2. Topol, E.J. High-Performance Medicine: The Convergence of Human and Artificial Intelligence. *Nat. Med.* **2019**, *25*, 44–56, doi:https://doi.org/10.1038/s41591-018-0300-7.
3. Hamet, P.; Tremblay, J. Artificial Intelligence in Medicine. *Metabolism* **2017**, *69*, S36–S40.
4. Newell, A.; Shaw, J.C.; Simon, H.A. Elements of a Theory of Human Problem Solving. *Psychol. Rev.* **1958**, *65*, 151.
5. Samuel, A.L. Some Studies in Machine Learning Using the Game of Checkers. *IBM J. Res. Dev.* **1959**, *3*, 210–229.
6. Warner, H.R.; Toronto, A.F.; Veasey, L.G.; Stephenson, R. A Mathematical Approach to Medical Diagnosis: Application to Congenital Heart Disease. *Jama* **1961**, *177*, 177–183.
7. Weizenbaum, J. ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine. *Commun. ACM* **1966**, *9*, 36–45.
8. De Dombal, F.T.; Leaper, D.J.; Staniland, J.R.; McCann, A.P.; Horrocks, J.C. Computer-Aided Diagnosis of Acute Abdominal Pain. *Br Med J* **1972**, *2*, 9–13.
9. Szolovits, P. Artificial Intelligence in Medical Diagnosis. *Ann. Intern. Med.* **1988**, *108*, 80, doi:10.7326/0003-4819-108-1-80.
10. Castelvecchi, D. Can We Open the Black Box of AI? *Nat. News* **2016**, *538*, 20, doi:10.1038/538020a.
11. Finlayson, S.G.; Bowers, J.D.; Ito, J.; Zittrain, J.L.; Beam, A.L.; Kohane, I.S. Adversarial Attacks on Medical Machine Learning. *Science* **2019**, *363*, 1287–1289.
12. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet Classification with Deep Convolutional Neural Networks. In Proceedings of the Advances in neural information processing systems; 2012; pp. 1097–1105.
13. Marcus, G. Deep Learning: A Critical Appraisal. *ArXiv Prepr. ArXiv180100631* **2018**.
14. Deo Rahul C. Machine Learning in Medicine. *Circulation* **2015**, *132*, 1920–1930, doi:10.1161/CIRCULATIONAHA.115.001593.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202012.0313.v2

Peer-reviewed version available at *Symmetry* **2021**, *13*, 102; doi:10.3390/sym13010102

15. Yu, K.-H.; Berry, G.J.; Rubin, D.L.; Re, C.; Altman, R.B.; Snyder, M. Association of Omics Features with Histopathology Patterns in Lung Adenocarcinoma. *Cell Syst.* **2017**, *5*, 620–627.
16. Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M.C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *Jama* **2016**, *316*, 2402–2410.
17. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. Chestx-Ray8: Hospital-Scale Chest x-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition; 2017; pp. 2097–2106.
18. Strodthoff, N.; Strodthoff, C. Detecting and Interpreting Myocardial Infarction Using Fully Convolutional Neural Networks. *Physiol. Meas.* **2019**, *40*, 015001.
19. Bejnordi, B.E.; Veta, M.; Van Diest, P.J.; Van Ginneken, B.; Karssemeijer, N.; Litjens, G.; Van Der Laak, J.A.; Hermsen, M.; Manson, Q.F.; Balkenhol, M. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women with Breast Cancer. *Jama* **2017**, *318*, 2199–2210.
20. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature* **2017**, *542*, 115–118.
21. Yu, K.-H.; Beam, A.L.; Kohane, I.S. Artificial Intelligence in Healthcare. *Nat. Biomed. Eng.* **2018**, *2*, 719–731.
22. Gerke, S.; Minssen, T.; Cohen, I.G. Ethical and Legal Challenges of Artificial Intelligence-Driven Health Care. *Forthcom. Artif. Intell. Healthc. 1st Ed. Adam Bohr Kaveh Memarzadeh Eds* **2020**.
23. Ntoutsi, E.; Fafalios, P.; Gadiraju, U.; Iosifidis, V.; Nejdl, W.; Vidal, M.-E.; Ruggieri, S.; Turini, F.; Papadopoulos, S.; Krasanakis, E.; et al. Bias in Data-Driven Artificial Intelligence Systems-An Introductory Survey. *Wiley Interdiscip. Rev.-Data Min. Knowl. Discov.* **2019**, e1356, doi:10.1002/widm.1356.
24. Vandewiele, G.; De Backere, F.; Lannoye, K.; Vanden Berghe, M.; Janssens, O.; Van Hoecke, S.; Keereman, V.; Paemeleire, K.; Ongenae, F.; De Turck, F. A Decision Support System to Follow up and Diagnose Primary Headache Patients Using Semantically Enriched Data. *BMC Med. Inform. Decis. Mak.* **2018**, *18*, 98, doi:10.1186/s12911-018-0679-6.
25. Kwon, B.C.; Choi, M.-J.; Kim, J.T.; Choi, E.; Kim, Y.B.; Kwon, S.; Sun, J.; Choo, J. RetainVis: Visual Analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records. *IEEE Trans. Vis. Comput. Graph.* **2018**, *25*, 299–309, doi:10.1109/TVCG.2018.2865027.
26. De Fauw, J.; Ledsam, J.R.; Romera-Paredes, B.; Nikolov, S.; Tomasev, N.; Blackwell, S.; Askham, H.; Glorot, X.; O'Donoghue, B.; Visentin, D. Clinically Applicable Deep Learning for Diagnosis and Referral in Retinal Disease. *Nat. Med.* **2018**, *24*, 1342–1350.
27. Ting, D.S.W.; Pasquale, L.R.; Peng, L.; Campbell, J.P.; Lee, A.Y.; Raman, R.; Tan, G.S.W.; Schmetterer, L.; Keane, P.A.; Wong, T.Y. Artificial Intelligence and Deep Learning in Ophthalmology. *Br. J. Ophthalmol.* **2019**, *103*, 167–175, doi:10.1136/bjophthalmol-2018-313173.
28. Bleicher, A. Demystifying the Black Box That Is AI Available online: https://www.scientificamerican.com/article/demystifying-the-black-box-that-is-ai/ (accessed on 4 March 2020).
29. Heaven, W.D. Why Asking an AI to Explain Itself Can Make Things Worse Available online: https://www.technologyreview.com/s/615110/why-asking-an-ai-to-explain-itself-can-make-things-worse/ (accessed on 4 March 2020).
30. Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; Vinyals, O. Understanding Deep Learning Requires Rethinking Generalization. *ArXiv Prepr. ArXiv161103530* **2016**.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202012.0313.v2

Peer-reviewed version available at Symmetry 2021, 13, 102; doi:10.3390/sym13010102

31. Schemelzer, R. Understanding Explainable AI Available online: https://www.forbes.com/sites/cognitiveworld/2019/07/23/understanding-explainable-ai/#406c97957c9e (accessed on 10 April 2020).
32. London, A.J. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Cent. Rep.* **2019**, *49*, 15–21, doi:10.1002/hast.973.
33. Wang, F.; Kaushal, R.; Khullar, D. Should Health Care Demand Interpretable Artificial Intelligence or Accept "Black Box" Medicine? *Ann. Intern. Med.* **2019**, *172*, 59–60, doi:10.7326/M19-2548.
34. Buehler, M.; Iagnemma, K.; Singh, S. *The DARPA Urban Challenge: Autonomous Vehicles in City Traffic*; springer, 2009; Vol. 56;.
35. Wang, X.; Bisantz, A.M.; Bolton, M.L.; Cavuoto, L.; Chandola, V. Explaining Supervised Learning Models: A Preliminary Study on Binary Classifiers. *Ergon. Des.* **2020**, *28*, 20–26, doi:10.1177/1064804620901641.
36. Teo, E.A.L.; Ling, F.Y.Y. Developing a Model to Measure the Effectiveness of Safety Management Systems of Construction Sites. *Build. Environ.* **2006**, *41*, 1584–1592.
37. Fernández-Muñiz, B.; Montes-Peon, J.M.; Vazquez-Ordas, C.J. Safety Management System: Development and Validation of a Multidimensional Scale. *J. Loss Prev. Process Ind.* **2007**, *20*, 52–68.
38. Liberati, A.; Altman, D.G.; Tetzlaff, J.; Mulrow, C.; Gøtzsche, P.C.; Ioannidis, J.P.A.; Clarke, M.; Devereaux, P.J.; Kleijnen, J.; Moher, D. The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *PLOS Med.* **2009**, *6*, e1000100, doi:10.1371/journal.pmed.1000100.
39. Davahli, M.R.; Karwowski, W.; Gutierrez, E.; Fiok, K.; Wróbel, G.; Taiar, R.; Ahram, T. Identification and Prediction of Human Behavior through Mining of Unstructured Textual Data. *Symmetry* **2020**, *12*, 1902, doi:10.3390/sym12111902.
40. Qu, S.Q.; Dumay, J. The Qualitative Research Interview. *Qual. Res. Account. Manag.* **2011**.
41. DiCicco-Bloom, B.; Crabtree, B.F. The Qualitative Research Interview. *Med. Educ.* **2006**, *40*, 314–321.
42. HAWKIRB Studies That Are Not Human Subjects Research Available online: https://hso.research.uiowa.edu/studies-are-not-human-subjects-research (accessed on 27 December 2020).
43. Kendall's W in Wikipedia Kendall's W. *Wikipedia* 2020.
44. Legendre, P. Species Associations: The Kendall Coefficient of Concordance Revisited. *J. Agric. Biol. Environ. Stat.* **2005**, *10*, 226.
45. Ćwiklicki, M.; Klich, J.; Chen, J. The Adaptiveness of the Healthcare System to the Fourth Industrial Revolution: A Preliminary Analysis. *Futures* **2020**, *122*, 102602.
46. Hale, A.R.; Baram, M.S. *Safety Management: The Challenge of Change*; Pergamon Oxford, 1998;
47. Matheny, M.E.; Whicher, D.; Israni, S.T. Artificial Intelligence in Health Care: A Report From the National Academy of Medicine. *Jama* **2020**, *323*, 509–510.
48. Zhu, Q.; Jiang, X.; Zhu, Q.; Pan, M.; He, T. Graph Embedding Deep Learning Guides Microbial Biomarkers' Identification. *Front. Genet.* **2019**, *10*, 1182, doi:10.3389/fgene.2019.01182.
49. Challen, R.; Denny, J.; Pitt, M.; Gompels, L.; Edwards, T.; Tsaneva-Atanasova, K. Artificial Intelligence, Bias and Clinical Safety. *BMJ Qual Saf* **2019**, *28*, 231–237.
50. Kelly, C.J.; Karthikesalingam, A.; Suleyman, M.; Corrado, G.; King, D. Key Challenges for Delivering Clinical Impact with Artificial Intelligence. *BMC Med.* **2019**, *17*, 195, doi:10.1186/s12916-019-1426-2.
51. Rose, S. Machine Learning for Prediction in Electronic Health Data. *JAMA Netw. Open* **2018**, *1*, e181404–e181404.
52. U.S. Food and Drug Administration; Administration, D. *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)-Discussion Paper*; Discussion paper and request for feedback, 2019;

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202012.0313.v2

Peer-reviewed version available at *Symmetry* **2021**, *13*, 102; doi:10.3390/sym13010102

53. Shah, P.; Kendall, F.; Khozin, S.; Goosen, R.; Hu, J.; Laramie, J.; Ringel, M.; Schork, N. Artificial Intelligence and Machine Learning in Clinical Development: A Translational Perspective. *NPJ Digit. Med.* **2019**, *2*, 1–5.
54. Nordling, L. A Fairer Way Forward for AI in Health Care. *Nature* **2019**, *573*, S103–S105, doi:10.1038/d41586-019-02872-2.
55. Stewart, E. Self-Driving Cars Have to Be Safer than Regular Cars. The Question Is How Much. Available online: https://www.vox.com/recode/2019/5/17/18564501/self-driving-car-morals-safety-tesla-waymo (accessed on 7 March 2020).
56. Golden, J.A. Deep Learning Algorithms for Detection of Lymph Node Metastases from Breast Cancer: Helping Artificial Intelligence Be Seen. *Jama* **2017**, *318*, 2184–2186.
57. Amershi, S.; Weld, D.; Vorvoreanu, M.; Fourney, A.; Nushi, B.; Collisson, P.; Suh, J.; Iqbal, S.; Bennett, P.N.; Inkpen, K. Guidelines for Human-AI Interaction. In Proceedings of the Proceedings of the 2019 chi conference on human factors in computing systems; 2019; pp. 1–13.
58. Salazar, J.W.; Redberg, R.F. Leading the Call for Reform of Medical Device Safety Surveillance. *JAMA Intern. Med.* **2020**, *180*, 179–180.
59. Ventola, C.L. Challenges in Evaluating and Standardizing Medical Devices in Health Care Facilities. *Pharm. Ther.* **2008**, *33*, 348.
60. Wang, L.; Haghighi, A. Combined Strength of Holons, Agents and Function Blocks in Cyber-Physical Systems. *J. Manuf. Syst.* **2016**, *40*, 25–34.
61. Callahan, A.; Fries, J.A.; Ré, C.; Huddleston, J.I.; Giori, N.J.; Delp, S.; Shah, N.H. Medical Device Surveillance with Electronic Health Records. *NPJ Digit. Med.* **2019**, *2*, 1–10.
62. Forcier, M.B.; Gallois, H.; Mullan, S.; Joly, Y. Integrating Artificial Intelligence into Health Care through Data Access: Can the GDPR Act as a Beacon for Policymakers? *J. Law Biosci.* **2019**, *6*, 317.
63. Westerheide, F. The Artificial Intelligence Industry and Global Challenges Available online: https://www.forbes.com/sites/cognitiveworld/2019/11/27/the-artificial-intelligence-industry-and-global-challenges/ (accessed on 9 March 2020).
64. Nicola, S.; Behrmann, E.; Mawad, M. *It's a Good Thing Europe's Autonomous Car Testing Is Slow.[Online] Bloomberg*; 2018;
65. Price, W.N.; Cohen, I.G. Privacy in the Age of Medical Big Data. *Nat. Med.* **2019**, *25*, 37–43, doi:10.1038/s41591-018-0272-7.
66. Wenyan, W. China Is Waking up to Data Protection and Privacy. Here's Why That Matters Available online: https://www.weforum.org/agenda/2019/11/china-data-privacy-laws-guideline/ (accessed on 29 February 2020).
67. Lindsey, N. China's Privacy Challenges with AI and Mobile Apps Available online: https://www.cpomagazine.com/data-privacy/chinas-privacy-challenges-with-ai-and-mobile-apps/ (accessed on 29 February 2020).
68. O'Meara, S. Will China Lead the World in AI by 2030? *Nature* **2019**, *572*, 427–428, doi:10.1038/d41586-019-02360-7.
69. US Department of Health and Human Services *Software as a Medical Device (SAMD): Clinical Evaluation. Guidance for Industry and Food and Drug Administration Staff. 2017*; 2017;
70. U.S. Food and Drug Administration; Administration, D. Digital Health Innovation Action Plan. *Digit. Health Program* **2017**.
71. Xu, H.; Caramanis, C.; Mannor, S. Sparse Algorithms Are Not Stable: A No-Free-Lunch Theorem. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 187–193.
72. Sun, W. Stability of Machine Learning Algorithms. *Open Access Diss. 563* **2015**, doi:https://docs.lib.purdue.edu/open_access_dissertations/563/.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202012.0313.v2

Peer-reviewed version available at *Symmetry* **2021**, *13*, 102; doi:10.3390/sym13010102

73. Cheatham, B.; Javanmardian, K.; Samandari, H. Confronting the Risks of Artificial Intelligence. *McKinsey Q.* **2019**, 1–9.
74. Langlotz, C.P.; Allen, B.; Erickson, B.J.; Kalpathy-Cramer, J.; Bigelow, K.; Cook, T.S.; Flanders, A.E.; Lungren, M.P.; Mendelson, D.S.; Rudie, J.D.; et al. A Roadmap for Foundational Research on Artificial Intelligence in Medical Imaging: From the 2018 NIH/RSNA/ACR/The Academy Workshop. *Radiology* **2019**, *291*, 781–791, doi:10.1148/radiol.2019190613.
75. Decencière, E.; Zhang, X.; Cazuguel, G.; Lay, B.; Cochener, B.; Trone, C.; Gain, P.; Ordonez, R.; Massin, P.; Erginay, A. Feedback on a Publicly Distributed Image Database: The Messidor Database. *Image Anal. Stereol.* **2014**, *33*, 231–234.
76. Yu, K.-H.; Zhang, C.; Berry, G.J.; Altman, R.B.; Ré, C.; Rubin, D.L.; Snyder, M. Predicting Non-Small Cell Lung Cancer Prognosis by Fully Automated Microscopic Pathology Image Features. *Nat. Commun.* **2016**, *7*, 12474.
77. Bhagwat, N.; Viviano, J.D.; Voineskos, A.N.; Chakravarty, M.M.; Initiative, A.D.N. Modeling and Prediction of Clinical Symptom Trajectories in Alzheimer's Disease Using Longitudinal Data. *PLoS Comput. Biol.* **2018**, *14*, e1006376.
78. Gianfrancesco, M.A.; Tamang, S.; Yazdany, J.; Schmajuk, G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Intern. Med.* **2018**, *178*, 1544–1547.
79. Kocheturov, A.; Pardalos, P.M.; Karakitsiou, A. Massive Datasets and Machine Learning for Computational Biomedicine: Trends and Challenges. *Ann. Oper. Res.* **2019**, *276*, 5–34.
80. Li, Z.; Wang, C.; Han, M.; Xue, Y.; Wei, W.; Li, L.-J.; Fei-Fei, L. Thoracic Disease Identification and Localization with Limited Supervision. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018; pp. 8290–8299.
81. Nam, J.G.; Park, S.; Hwang, E.J.; Lee, J.H.; Jin, K.-N.; Lim, K.Y.; Vu, T.H.; Sohn, J.H.; Hwang, S.; Goo, J.M. Development and Validation of Deep Learning–Based Automatic Detection Algorithm for Malignant Pulmonary Nodules on Chest Radiographs. *Radiology* **2019**, *290*, 218–228.
82. Steiner, D.F.; MacDonald, R.; Liu, Y.; Truszkowski, P.; Hipp, J.D.; Gammage, C.; Thng, F.; Peng, L.; Stumpe, M.C. Impact of Deep Learning Assistance on the Histopathologic Review of Lymph Nodes for Metastatic Breast Cancer. *Am. J. Surg. Pathol.* **2018**, *42*, 1636.
83. Chilamkurthy, S.; Ghosh, R.; Tanamala, S.; Biviji, M.; Campeau, N.G.; Venugopal, V.K.; Mahajan, V.; Rao, P.; Warier, P. Deep Learning Algorithms for Detection of Critical Findings in Head CT Scans: A Retrospective Study. *The Lancet* **2018**, *392*, 2388–2396.
84. Haenssle, H.A.; Fink, C.; Schneiderbauer, R.; Toberer, F.; Buhl, T.; Blum, A.; Kalloo, A.; Hassen, A.B.H.; Thomas, L.; Enk, A. Man against Machine: Diagnostic Performance of a Deep Learning Convolutional Neural Network for Dermoscopic Melanoma Recognition in Comparison to 58 Dermatologists. *Ann. Oncol.* **2018**, *29*, 1836–1842.
85. Yasaka, K.; Akai, H.; Abe, O.; Kiryu, S. Deep Learning with Convolutional Neural Network for Differentiation of Liver Masses at Dynamic Contrast-Enhanced CT: A Preliminary Study. *Radiology* **2018**, *286*, 887–896.
86. Kermany, D.S.; Goldbaum, M.; Cai, W.; Valentim, C.C.; Liang, H.; Baxter, S.L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell* **2018**, *172*, 1122–1131.
87. Jaremko, J.L.; Azar, M.; Bromwich, R.; Lum, A.; Alicia Cheong, L.H.; Gibert, M.; Laviolette, F.; Gray, B.; Reinhold, C.; Cicero, M.; et al. Canadian Association of Radiologists White Paper on Ethical and Legal Issues Related to Artificial Intelligence in Radiology. *Can. Assoc. Radiol. J. J. Assoc. Can. Radiol.* **2019**, *70*, 107–118, doi:10.1016/j.carj.2019.03.001.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202012.0313.v2

Peer-reviewed version available at *Symmetry* **2021**, *13*, 102; doi:10.3390/sym13010102

88. Patel, N.M.; Michelini, V.V.; Snell, J.M.; Balu, S.; Hoyle, A.P.; Parker, J.S.; Hayward, M.C.; Eberhard, D.A.; Salazar, A.H.; McNeillie, P. Enhancing Next-Generation Sequencing-Guided Cancer Care through Cognitive Computing. *The oncologist* **2018**, *23*, 179.
89. CBINSIGHTS Google Healthcare With AI l CB Insights Available online: https://www.cbinsights.com/research/report/google-strategy-healthcare/ (accessed on 7 March 2020).
90. Miotto, R.; Li, L.; Kidd, B.A.; Dudley, J.T. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci. Rep.* **2016**, *6*, 1–10.
91. Camacho, D.M.; Collins, K.M.; Powers, R.K.; Costello, J.C.; Collins, J.J. Next-Generation Machine Learning for Biological Networks. *Cell* **2018**, *173*, 1581–1592.
92. Hill, J. Simulation: The Bedrock of AI Available online: https://medium.com/simudyne/simulation-the-bedrock-of-ai-12153eaf7971 (accessed on 8 March 2020).
93. Chawla, V. How Training AI Models In Simulated Environments Is Helping Researchers. *Anal. India Mag.* 2019.
94. O'Kane, S. Tesla and Waymo Are Taking Wildly Different Paths to Creating Self-Driving Cars Available online: https://www.theverge.com/transportation/2018/4/19/17204044/tesla-waymo-self-driving-car-data-simulation (accessed on 2 March 2020).
95. Upton, R. Artificial Intelligence's Need for Health Data – Finding An Ethical Balance. *Hit Consult.* 2019.
96. Wang, F.; Preininger, A. AI in Health: State of the Art, Challenges, and Future Directions. *Yearb. Med. Inform.* **2019**, *28*, 016–026.
97. Hsu, J.; Spectrum AI Could Make Detecting Autism Easier Available online: https://www.spectrumnews.org/features/deep-dive/can-computer-diagnose-autism/ (accessed on 19 February 2020).
98. Christian, J.; Dasgupta, N.; Jordan, M.; Juneja, M.; Nilsen, W.; Reites, J. Digital Health and Patient Registries: Today, Tomorrow, and the Future. In *21st Century Patient Registries: Registries for Evaluating Patient Outcomes: A User's Guide: 3rd Edition, Addendum [Internet]*; Agency for Healthcare Research and Quality (US), 2018.
99. Sayeed, R.; Gottlieb, D.; Mandl, K.D. SMART Markers: Collecting Patient-Generated Health Data as a Standardized Property of Health Information Technology. *Npj Digit. Med.* **2020**, *3*, 1–8.
100. U.S. Food and Drug Administration Medical Device Data Systems, Medical Image Storage Devices, and Medical Image Communications Devices Available online: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/medical-device-data-systems-medical-image-storage-devices-and-medical-image-communications-devices (accessed on 11 April 2020).
101. Rong, G.; Mendez, A.; Assi, E.B.; Zhao, B.; Sawan, M. Artificial Intelligence in Healthcare: Review and Prediction Case Studies. *Engineering* **2020**.
102. Machanick, P. Approaches to Addressing the Memory Wall. *Sch. IT Electr. Eng. Univ. Qld.* **2002**.
103. Devalla, S.K.; Liang, Z.; Pham, T.H.; Boote, C.; Strouthidis, N.G.; Thiery, A.H.; Girard, M.J.A. Glaucoma Management in the Era of Artificial Intelligence. *Br. J. Ophthalmol.* **2020**, *104*, 301–311, doi:10.1136/bjophthalmol-2019-315016.