|  |  |  |
| --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION**TELECOMMUNICATIONSTANDARDIZATION SECTOR**STUDY PERIOD 2017-2020 | FG-AI4H-J-037 |
| **ITU-T Focus Group on AI for Health** |
| **Original: English** |
| **WG(s):** | DASH | E-meeting, 30 September – 2 October 2020 |
| **DOCUMENT** |
| **Source:** | ETH Zurich |
| **Title:** | Practical challenges of artificial intelligence in healthcare: building the evaluation framework for organizations |
| **Purpose:** | Discussion |
| **Contact:** | Agnieszka SitarskaSwitzerland | Tel: +789527771Email: Sitarska.aga@gmail.com |

|  |  |
| --- | --- |
| **Abstract:** | This document describes the practical aspects of AI-based systems and tools that healthcare organizations should assess before deciding on implementation. The criteria were uncovered through 20 interviews with experienced business practitioners, complemented by the insights from recent literature as well as learnings from the empirical evaluation project. The analysis gave ground to a thesis which concluded the MAS MTEC study program at ETH. The thesis proposed an evaluation framework for AI-based systems, which may serve as a practical guideline to business generalists from the healthcare industry. The framework starts with a set of preparatory activities to be conducted by the adopting organization. The internal assessment is followed by six clusters of external evaluation criteria: compliance, capabilities of the provider, technology and training data, design and user experience, verification and clinical proof, and lastly, integrability. |

There is an ever-growing need for digital innovation in healthcare due to increasing demand for care as well as the abundance of data which is now available for analysis in a digital form. In the current decade, the governments of developed countries have been incentivizing digitization via various legal acts which impact healthcare actors and technology providers alike.

Despite the growing enthusiasm, healthcare organizations are often hesitant in implementing innovation within their ecosystem, particularly if such innovation could affect the patients. Especially the actors with the more traditional structure do not seem equipped to carry out a thorough and systematic assessment of advanced technologies and providers. Thus, selecting a solution which would be the best fit for their needs proves very difficult.

Over the course of my MAS thesis project, the author investigated the challenges related to the evaluation and selection of tools based on AI and intended for healthcare. The goal of the thesis was to create a practical framework which would provide a useful guide to business generalists who are often in charge of such evaluation efforts.

The project started off by exploring the most recent literature on evaluation and implementation of AI in healthcare (publications released in 2019-2020). In addition to that, insights were gathered from 20 business practitioners coming from both the client and the provider side. The interviewees represented various profiles: data scientists, innovation officers, medical and compliance professionals, experts in IT and digital services, and finally investors who face the dynamics of the early innovation markets on a daily basis.

On behalf of a sponsor company, an empirical process was carried out of scouting, preselection and evaluation of prospective providers of a patient-facing digital service based on AI. This experience allowed revision of the framework criteria and adjustment of its course to the empirical needs of an organizational setting.

The gathered insights enabled the discrimination of 7 clusters of evaluation criteria. The first one consists of preparatory assessments which the organization should consider internally, at the very start of the innovation initiative. The following 6 clusters comprise evaluation criteria that are to be used when a prospective provider (one or more) has been identified. The general structure of the framework is presented in Figure 1.



Figure 1. General layout of the framework

The structure of the general layout of the framework in Figure 1 consists of two main parts: (A) internal groundwork, which an evaluating organization is advised to complete at the start of the project; (B) Evaluation criteria, which consists of six distinct categories: Provider (with sub-section specific to startups), Technology and training data, Design and user experience (UX), Verification and Clinical proof, Compliance and Integratability.

The following part of this document presents the details of each of the 7 clusters.

# Internal groundwork

It is important at the very start of the selection process that project sponsors define the **strategic aim** that the company intends to achieve by implementing the solution as well as precise the expectations from the tool. This step is very important, since particular goals may alter the course of evaluation, prioritizing certain metrics over other as well as determine the integration approach.

Clarifying the scope and expectations shall be complemented by defining the high-level **specifications** of the target solution, which will serve as exclusion criteria for the initial market screening. The core criteria shall be in line with the company strategy. At this stage the organization should assess its **internal IT infrastructure and context**, along with specific processes which will be affected by the innovation. Selecting the tool that is compatible with the specific organizational context helps to ensure feasibility of the implementation further on.

Knowing the scope, the strategic aim and the core specifications allows assessing the **regulatory and legal landscape**. If any authorities need to be notified about the planned new internal solution or a client-facing digital service, it is best to find out early what compliance requirements need to be fulfilled before obtaining an approval.

While screening the innovation market for suitable solutions, it is worthwhile to consider the **maturity and structure** of the particular **market niche**. If the market is immature, one should expect a higher number of small vendors, probably startup companies. Analyzing the competitive environment helps get the feeling of what strategies and features seem to be effective in deployment, which can be inferred from commercial success. It helps to better understand the nature of the sought innovation.

In order to improve coordination, the organization should establish a **network of experts** across departments who can provide their input in various stages of the analysis. The evaluation topics may be split for analysis among the technical, medical and business practitioners. However, all learnings should be shared within the whole group, to ensure the overall progress in the process and to eliminate redundant iterations with the provider.

The importance of **communication** goes beyond sharing information within the team of experts. It is important that evaluators explain their selection process to the **decision makers** early on as well, ensuring their commitment to the project as well as constructive and fact-based discussion at the end of evaluation.

Having considered all the internal aspects, the evaluators may proceed to scrutinizing the external aspects against the criteria listed in the following section.

# Evaluation criteria

This part of the framework presents the 6 external-facing evaluation clusters, which may serve as a guide in evaluating external providers. Each subsection presents an overview of one clusters as well as required resources that need to be involved in the assessment.

Table 1 – Framework external-facing evaluation clusters

|  |  |
| --- | --- |
| **#** | **Evaluation category** |
| 1 | Compliance |
| 2 | Capabilities of the provider |
| 3 | Technology and training data |
| 4 | Design and user experience |
| 5 | Verification and clinical proof |
| 6 | Integratability |

## 2.1 Compliance

As in every innovation project and particularly when personal and health data are involved, the compliance aspects are critical in decision making. Figure 2 presents the layout of the identified aspects, as expressed by the interviewed practitioners. This part of the evaluation process should be handled by the procurement department, complemented by regulatory affairs and legal expertise.



Figure 2. Evaluation aspects in scope of the Compliance criterion

The evaluators should start with controlling if the vendor has all necessary **certifications** as required for the specific type of solution. In order to determine which regulations apply, it is important to define in detail the **scope of entered, processed and output data**. Further, the evaluators must find out in what **country** the data is stored and if it is transferred to any other location for training or any other purpose. It is well perceived when the vendor has successfully passed an **audit** by the local data protection authorities. Additionally, the provider should have a defined and formalized **QA process** for the post-deployment surveillance of the tool.

The other data-related aspects of compliance which should be ensured at the evaluation stage is the soundness of **data** **security measures** that must be part of the solution. The solution should collect only as much data as necessary to produce an output and should process them in an encrypted and anonymized form. The vendor should have a sound approach to communication with data owners in order to obtain their **informed consent** for collecting, processing and storing data. The informed consent is required not only for the operational data, but also the sets used in the development of the tool.

Since the **regulatory landscape and standards** surrounding AI in healthcare is only being shaped right now, it is important that evaluators as well as vendors remain vigilant and monitor the progress.

## 2.2 Capabilities of the provider

This section presents a synopsis of what the interviewees considered to be the most important qualities in the providers of the evaluated solutions. There are a few specific aspects which pertain to startup companies and should be considered in addition to the general scope. That stage of analysis requires input from the procurement department. Data science expertise is highly important as well, as it helps to understand the competences of a vendor as well as their ability to provide ongoing support over the course of the product lifecycle.

Figure 3 presents the layout of the respective evaluation aspects, with a dedicated subsection for startup companies which complements the general part. The evaluation aspects in scope are those of of the Provider category. There are a few aspects that are applicable specifically to startup vendors.



Figure 3. Capabilities of the provider

One of the most important aspects to check is the **AI expertise** in the team. Evaluators should be interested in particular in the profile of the person heading the AI efforts of the vendor. The relevant background and experience of that leader gives evaluators confidence in the quality of the investigated solution. It is considered beneficial if a vendor maintains connections to research innovation hubs, which gives it permanent access to the scarce **machine learning talent**.

**Domain expertise** is also considered essential. As such, it should be present in the management of the evaluated enterprise. In the case of most solutions related to healthcare, domain expertise would be provided by medical doctors, ideally with practical, clinical experience in their resume.

Furthermore, the provider should have a strong expertise in designing **user experience (UX)**. Ideally, there would be a dedicated specialist in their ranks who permanently supervises the UX side of the product.

It is preferred when a vendor is able to provide support on the **regulatory and legal** aspects. This stage of integration was often said to be a challenge. Given the uncertainty that surrounds AI-based solutions for healthcare, adopting organizations may strongly benefit from the professional support. If a vendor has regulatory and legal personnel, it gives clients confidence that the development process of their chosen solution was sound, resulting in a truly compliant end-product.

It is beneficial when a vendor has a dedicated **quality assurance (QA) team** who supports in continuous monitoring of the tool and any necessary improvements. Maintaining an established communication channel, e.g. **helpdesk** for problems and feedback, is beneficial as well.

Not only expertise but also **practical experience** in deploying AI solutions is considered critically important by the interviewees. The experience should be from **outside of academia**, preferably in the **target industry**. Evaluators value when a vendor has ongoing partnerships or clients in their sector, as it guarantees a sound understanding of the specific needs and challenges.

Good **reputation** of the provider among its existing clients and partners plays an important role in the evaluation as well. Moreover, experience in managing organizational, **“business to business” (B2B) clients** was said to be beneficial, since it is known to have particular dynamics and requires a specific approach.

Since the ongoing and long-term support from the vendor should be expected for solutions based on AI, **stability** of the company should be looked into as well. That aspect is particularly important when a vendor is a young enterprise, since startups are known as high-risk ventures.

A few further evaluation aspects were expressed in reference to **startup companies** in particular and do not apply to established providers. In terms of the age of a startup, **4-5 years** should be expected as a minimum, since this is how long finding the optimal market fit seems to take. In terms of stability, the coherence of the **founding team** should be looked into, as well as the **financing stage** of the enterprise. Startups which reach funding round B or further, should be favored over ventures at the beginning of their journey, because they have been scrutinized by a number of investors who bet their money on the soundness of the overall solution. With the same rationale, it is desirable to see prominent names among **investors** backing the provider. In case of startups, the **vision** of the company is important as well, because it gives an approximation of the direction how the product and the partnership may **develop over time**.

## 2.3 Technology and training data

Most of the interviewees were of the opinion that scrutinizing the type of AI model does not play a prominent part in their evaluation approach. More important is the suitability of the approach to a defined use case and the degree to which its working can be explained. This section presents the overview of evaluation aspects related to the technology and training approach. Figure 4 presents the layout of the identified relevant aspects. That part of the evaluation would benefit from involving data science as well as IT expertise.



Figure 4. Evaluation aspects in scope of the technology and training data criterion

When looking at the **algorithm** at the core of the solution, evaluators want to make sure that it is state-of-the-art and will remain so in the foreseeable future. The type of the model must be **applicable** to the particular intended use and setting. It is considered beneficial if the vendor used established and sound AI frameworks as building blocks to create the solution. However, it is important that the tool has **proprietary code** in its composition as well, that adds differentiating functionalities to the solution.

Specifications of the solution should be gathered in a comprehensive **documentation package**, which should be made available to the evaluators. This prevents misunderstandings in the assessment process. In addition to that, the working of the evaluated solution should be made clear and should be **comprehensible** to all involved stakeholders, regardless of their background. Understanding the fundamentals is seen as a precondition for a successful adoption.

A lot of attention is given to the **training data**. **Adequate, well-structured and correctly annotated** training data are very important to the further performance of the AI-based solution. It is beneficial if the vendor had access to **unique datasets**, for example shared by partners and composed of real-world cases, very close to the operational setting. It is desirable to have the tool trained on **diverse datasets** in **sufficient quantity**.

The quality of **labels** in the training sets is important as well. Annotation of the training data should be performed by subject matter experts. The specialists should preferably have access to the final patient outcomes which constitute an objective reference, known as “ground truth”. Any **deficiencies** or **biases** in the training data should be disclosed (e.g. clinical datasets from one specific ethnical group) and reflected in the tool specifications.

In terms of **performance metrics**, the market standard ROC curve may serve as a starting point of the analysis. However, evaluators wish to see also the performance metrics **specific** to the deployment case (e.g. sensitivity, time to output). They are also interested to know if certain parameters can be adjusted to improve the fit of the solution to the operational setting.

It is important to know how the solution handles **varying quality of input data** and if it has any mechanisms in place to improve it. It is desirable if the tool is able to **scale** by including additional formats of input data. Also in case of changes in the operational setting, e.g. addition of new medical guidelines, the provider must be able to adjust the solution accordingly. Such adjustment often requires **re-training** the solution on new datasets. It is important to understand how this process works, since it may add additional consideration related to technology, compliance or other.

Finally, evaluators need to understand how the **auto-didactic capabilities** of AI are managed and **controlled** in the evaluated tool. In the sensitive and heavily regulated healthcare setting, the most desirable option is to keep updates and maintenance under human control.

## 2.4 Design and user experience

The interviewed professionals agreed that aspects related to solution design and user experience are key to the adoption and optimal performance of AI-based solutions. The identified evaluation aspects are presented in Figure 5. The expertise in digital services and IT was instrumental in uncovering those criteria and thus it may be beneficial to involve the respective departments at this stage of the evaluation process.



Figure 5. Evaluation aspects in scope of the design and user experience (UX) criterion

The assessment efforts may start with considering the overall **value proposition**. The previous section focused on its technology side, whereas this one considers it from the perspective of a defined user need. Also, the design of the **front-end** of an AI-based solution plays an important role, in particular when the evaluated tool is a digital service intended for the evaluator’s clients. In addition to front-end design, particular attention should be given to **user experience (UX)**: the user flow should be analysed in detail, to detect any potential inconsistencies. Evaluators need to ensure that the future users will be able to interact with the tool as intended.

An element of UX which may directly translate into the performance of the tool, is its approach to **capturing data**. The solution must use a balanced approach to elicit as much data as possible and **verify the input** on the spot, in a way that does not intimidate the user. Pleasant UX and **ease of navigation** should be maintained at all cost, to ensure adherence to the solution.

It is important that the end user understands the working of the tool as well as its outputs. In order to ensure that, the tool should demonstrate a **rationale** supporting its predictions. Such rationale could have the form of an approximated elaboration of how and exactly what parts of the input data were processed into the particular output. It the user does not understand the output, they should have the option to receive a **clarification** from the tool. This is particularly important in the case of diagnostic support tools, where users are expected to follow the recommendation given by the tool. In such a case it is important that the tool shows the **confidence level** along with the calculated output. Having this information, the user can take an informed and autonomous decision regarding the next steps. The confidence level should have a defined threshold below which the tool would enter the **failure mode**. Otherwise it may confuse the user with an advice which is not sound.

In addition to the usability and transparency aspects, there should be a **feedback loop** embedded in the tool design. This enables the tool to continually compare its output with real-world user outcomes and decisions, and make relevant adjustments to its prediction models. To take full advantage of the auto-didactic capabilities of AI, it is important that the users of the tool timely install all released **updates**. The updates must be compatible with the operating system of the end device and must remain functional amid continual updates of that system.

Given the novelty of AI-based solutions, the adopting company should require **continuous monitoring** of the use of the tool, its performance and in particular erroneous outputs. Such analytics are essential to measure the adoption as well as to make timely improvements where necessary. If the tool was deployed in a practical setting before, such analytics may be available for review. Apart from the various aspects of performance, it is beneficial if a vendor can demonstrate the **effectiveness** of the tool. Such insights can be derived either from past deployment cases or approximated.

## 2.5 Verification and clinical proof

Given the complex and data-driven nature of AI-based solutions, their quality can often only be judged by the output. This section presents the evaluation aspects related to verification, at the stage of the PoC as well as the following clinical validation for the specific intended purpose. Figure 6 shows the layout of the relevant aspects.



Figure 6. Evaluation aspects in scope of the verification and clinical proof criterion

The qualitativeresults of the proof of concept studies or the numerical values of the AUC demonstrated in the clinical validation are not sufficient to inform the evaluation. They must be analysed along with detailed **study protocols**, in order to understand the conditions and context of the study setup. At this stage of analysis, the expertise of data science professionals should be accompanied by a medical team, experienced in clinical evaluations.

A lot of attention is given to the **testing data** used as early as in the development process. The testing sets must be diverse and representative of the operational setting, however not too similar to the training data. It is best if the performance of the tool at the PoC or validation stage was demonstrated on **real-world cases**.

Some of the interviewed practitioners expressed that they feel more confident towards a solution which was part of **long-term research studies** carried out by reputable academic institutions. Ideally, the results of PoC studies are published in a peer-reviewed scientific journal.

When it comes to **performance metrics**, the ones defined and demonstrated by the vendor are a valuable starting point, ideally complemented by a live demonstration of the functionality of the tool. However, eventually the provider should be able to propose and demonstrate a performance metric that is specific to the target use case and the operational setting.

The interviewees repeatedly stated that in their practice they tend to request a test to be done on a sample of the **operational data**. That gives a good opportunity to verify the tool in the **target setting**, challenge it with **difficult or ambiguous cases** or even outright test its failure mode. Such test can often only be done at an advanced stage of evaluation, since it requires adjusting the tool to the target setting. The operational exercise presents an additional evaluation opportunity: allowing to verify the vendor’s claims on the **speed of the setup process**. In the interviewees experience, this often differs in practice to what providers promise at the beginning of the decision making process.

## 2.6 Integratability

This section contains the aspects related to the integration of the tool in the existing infrastructure. The criteria cover the context of the tool, including not only the surrounding IT systems but also the broad social context of stakeholders and user community. Figure 7 presents the layout of evaluation aspects related to the integratability.



Figure 7. Evaluation aspects in scope of the integratability criterion

Apart from the **compatibility** of the tool with the overall system of the adopting organization, the evaluators should ask about technical requirements to maintain the functionality of the tool. Some solutions may require expanding the adopter’s **IT infrastructure**, which is a serious cost factor that must be considered in the overall calculations.

The cost of integration should be considered not only in terms of the price of the license, but should also cover the time and resources. All those **cost drivers** shall be factored in when discriminating among competing solutions.

The solution needs to seamlessly **fit in the user journey**, as defined by the adopting organization. The **payment scheme** should be **accepted by the client** as well, and as close as possible to their **existing processes**.

In order to be attractive and feasible to a healthcare client, the AI-based product should have its established place within the relevant **regulatory landscape**. The path to regulatory approval should not be a challenge to an adopting organization.

**Summary**

The capabilities of the provider is the broadest among the six evaluation clusters. All the informants gave it a lot of attention during the interviews. The highly iterative dynamics of AI-based innovation requires very close collaboration of the adopting organization and the provider, seamless exchange of information and a transparent approach on both sides. That holds true not only at the stage of evaluation, but also after an implementation and practically throughout the lifecycle of a solution. This is why it is very important to select a vendor who can serve as a reliable partner over the long term.

The framework was built with the emphasis on a practical perspective, therefore it is suitable for business generalists for whom such evaluation projects are a new experience. It is very important to identify all relevant stakeholders at the stage of internal groundwork, so that working through the groups of criteria becomes the collective effort of all the relevant perspectives. The proposed approach does not impose any specific hierarchy or the order of analysis of the distinct groups of aspects. That gives evaluators a lot of flexibility in adjusting the approach to their own timelines, resources and priorities.

**Limitations**

Despite the versatility of the available business perspectives, interviews as a method of data collection did not allow for deep exploration of each identified aspect or implementation niche. Therefore, the final framework should be treated as a starting point or guidance in the evaluation efforts and not as an exhaustive procedure.

It must be noted that all the interviewed informants were affiliated with organizations who actively pursue innovations. The few startup representatives did share their experience from collaborating with hospitals, however, in general, the perspective of traditional providers of healthcare services was underrepresented in the collected material. Thus, as long as the resulting framework may be generalizable across early adopters of innovation, less innovative organizations may have additional criteria for the evaluated solutions. The framework may need to be revised by this type of adopters before it can be effectively applied in their setting.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_