MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220207

Artificial Intelligence and Clinical Decision Support Systems or Automated Interpreters: What Characteristics Are Expected by French General Practitioners?

Smaïn TABLA[a], Matthieu CALAFIORE[a,b], Bertrand LEGRAND[a,c], Axel DESCAMPS[a], Charlotte ANDRE[a], Michaël ROCHOY[a,d], Emmanuel CHAZARD[a]

[a] Univ. Lille, CHU Lille, ULR 2694 - METRICS, CERIM, Public health dept, F-59000 Lille, France
[b] General Practice, F-59150 Wattrelos, France
[c] General Practice, F-59200 Tourcoing, France
[d] General Practice, F-62230 Outreau, France

Abstract

Development of artificial intelligence (AI) modules should rely on technical progress, but also on users' needs. Our objective is to identify criteria that make a hypothetical AI module desirable for general practitioners (GPs). Method: random selection of 200 French GPs, and paper-based questionnaire. Results: the population was representative. GPs expect AI modules to diagnose or eliminate an urgent pathology for which they are not competent and for which specialists are not available. They also demand interoperability, automated electronic health record integration and facilitated information sharing. GPs would like AI modules to make them save time, simplify some procedures and delegate tasks to the secretary. They expect AI modules to allow them to associate the patient with the care, to reassure him or her, and to personalize the care. Interestingly, GPs would also rely on a machine to cut off abusive requests, such as work stoppages or certificates of convenience.

Keywords:
Artificial intelligence, computerized interpreter, automated diagnosis, usability.

Introduction

Intelligence denotes as a set of mental capacities for reasoning, problem solving and learning; an ability to integrate cognitive functions such as perception, attention, memory, language, or planning [1]. Artificial intelligence (AI) refers to the ability of a process to respond to environmental stimuli (or new data) in a manner whose results are comparable to that of human beings [2].

The first level of AI consists in setting up an inference engine connected to a knowledge base explicitly described by experts. This inference engine analyzes the data it receives and can, for example, issue a detailed message. This type of AI module is currently widely deployed in healthcare. The second level of AI takes advantage of machine learning techniques. A first step consists in automatically generating the rule base, generally in the form of predictive models, by using a learning dataset. The second step is then similar to the one described in level 1 AI. This second level takes advantage of deep learning methods.

In healthcare, AI is mainly used to support medical decision. More precisely, it can be used for disease screening or triage, diagnostic assistance, risk analysis, or assistance in setting up, adapting and monitoring treatment [3].

The applications of AI in healthcare are ubiquitous. We will cite a few examples here. In any case, it is important to keep in mind that the outputs proposed by AI modules are subject to error, and most of the time do not aim at replacing a health professional, but only at helping him/her, or saving time [3]. AI applications are numerous in imaging. They are also developing in ophthalmology for the screening and diagnosis of glaucoma [4] and diabetic retinopathy on fundus photographs. Applications also exist in oncology [5], in cardiology [5], in suicide prevention [6] or in the COVID-19 pandemic [7].

Some medical devices with artificial intelligence are also intended for use in general medicine. We can mention "intelligent" stethoscopes, which automatically classify the sounds heard during pulmonary auscultation [8] or cardiac auscultation [9], with the possibility of remote transmission. Automated electrocardiogram (EKG) interpreters have also invaded the daily life of general practitioners (GPs) [10]. In France in 2018 already, a quarter of French GPs were equipped with electrocardiographs with automated interpreters [10].

The development of health technologies follows a technology-driven process: researchers make a technique possible, evaluate it, and then an industrialist manages to raise funds to finance a product development. Usability evaluations of an AI module do exist [11], but they remain relatively rare.

It seems to us that the process is first technical, then financial, but does not necessarily stem from a user need. We have been asked to develop products that are technically feasible, but totally unrealistic in terms of workflow, such as the automated interpretation of an ultrasonographic image, when it is obvious that only an experienced sonographer is capable of capturing the image to be analyzed, and that this same person no longer needs an AI module to interpret it. Many technically mature processes may not find a market if they do not meet a need, or meet it in a non-useful or non-acceptable way.

Our hypothesis is that the acceptability criteria of an AI module should be clearly identified before seeking to develop it. Our objective is therefore to identify the domain-independent criteria that would make a hypothetical AI module desirable for GPs.

Methods

Design of the questionnaire

The questionnaire was defined by the authors. It was tested with 8 other GPs and iteratively improved. It includes a part allowing to describe the interviewed practitioner. The main part of the questionnaire consists of a set of 7-modality Likert scales. The main question is: "Imagine in the future a hypothetical device with a diagnostic or therapeutic decision support module (e.g., EKG or otoscope with computerized interpretation). This excludes web sites and medical biology sample analysis. For each of the following features, assuming it applies to this hypothetical device, would it encourage you to purchase such a device or not?" The items then proposed can be seen in table 2 of the results section. The 7 possible responses are for each item (for some statistical analyses, we will use the number in brackets):

- I would never buy it (-3)
- I am very discouraged (-2)
- It discourages me a little (-1)
- It has no impact (0)
- It encourages me a little (+1)
- It strongly encourages me (+2)
- I would definitely buy it (+3)

Intentionally, the term "artificial intelligence" is never used in the questionnaire, because from our previous experience its definition is not clearly known by practitioners, and this term is subject to fantasies.

Recruitment of participants

Participants were randomly drawn from a national directory of GPs. A paper letter with a pre-stamped return envelope was systematically sent to them. As the survey was strictly anonymous, all included GPs (even those who had responded) were called back by telephone after 2 weeks to improve the response rate.

Data analysis

Descriptive analyses were performed for each variable. The items evaluated using Likert scales were classified according to the mean numeric answer: "incentive items" from +3 to +0.7, "neutral items" from +0.7 to -0.7, and "deterrent items" from -0.7 to -3.

We tested the independence between each Likert scale (using its numeric value) and the participant's sex, the urban location of the practice, and the participant's age. Only statistically significant results were then reported.

Statistical analysis

Categorical variables were expressed in numbers and percentages. Quantitative variables were expressed as mean and standard deviation (SD) if the histogram revealed a symmetrical pattern distribution, and median first and third quartile (Q1, Q3) otherwise.

The independence between a quantitative variable and a categorical variable was tested using Welch's t-test. The independence between two quantitative variables was tested using Pearson's correlation coefficient nullity test.

Statistical tests were bilateral. The p values were considered significant at the 5% threshold.

Results

Analyzed questionnaires

Two hundred GPs were randomly selected. All of them were sent the questionnaire. No mail was returned in error. Among them, 139 (69.5%) answered the survey. Thirteen questionnaires (9.3%) were excluded because of poor quality (e.g., all the answers to Likert scales were identical). In total 126 questionnaires could be analyzed.

Participants' characteristics

There were 66 (52.4%) women, the average age was 47.8 years (SD=11.2). Men were significantly older than women (50.5 vs 45.4, p= 0.011). Among participants, 88 (71.0%) worked in a group practice, 32 (25.8%) worked alone, and 4 (3.2%) worked in another organization. The practice location was urban in 90 cases (71.4%), semi-rural in 30 cases (23.8%) and rural in 6 cases (4.8%). Participants were significantly younger in urban practices than in rural or semi-rural practices (respectively 45.5, 52.0 and 54.0 years, p=0.0006).

Table 1 reports the devices the GPs report having already used or had at disposal during consultations.

Table 1 – Devices with AI that GPs already used

Device	N (%)
EKG with computerized interpretation	46 (36.5%)
Connected blood pressure monitor	8 (6.3%)
(Semi-)automatic defibrillator	8 (6.3%)
Connected saturometer	3 (2.4%)
Connected bathroom scale	2 (1.6%)
Connected glucose meter	1 (0.8%)
Connected thermometer	1 (0.8%)
Connected smart stethoscope	0 (0.0%)
Dermatological pathology diagnosis	0 (0.0%)

Practitioners' attitude toward AI modules characteristics

Table 2 reports the 33 items that were evaluated using Likert scales. Items are sorted according to the average of the answers analyzed as numbers ranging from "-2 (I would never buy it)" to "+2 (I would definitely buy it)".

"Incentive items" (Table 2) relate to AI modules that:

- Enable to make or eliminate an urgent diagnosis
- Enable to make a diagnosis for which the GP is not competent, or for which no specialist is easily available
- Enable to register information in the electronic health record (EHR) or to send information to healthcare professionals
- Enable to simplify a procedure, save time, spend more time with the patient, or delegate tasks to the secretary
- Enable to predict a risk category or to perform personalized medicine
- Enable to involve the patient and reassures him/her
- Enable to cut short patients' abusive requests (e.g.: certificates, work stoppages, etc.)
- Train the GP him/herself

Table 2 –Results analysis of Likert scales: "Would this feature encourage you to, or discourage you from purchasing this module?"

Form order	Item	Mean	Conclusion	Preferred by
9	It quickly eliminates an urgent diagnosis	2.25	Incentive	-
10	It quickly makes an urgent diagnosis	2.19	Incentive	young (p=0.003)
21	It automatically adds information to the patient's electronic records	1.86	Incentive	young (p=0.01)
20	It allows me to quickly share an examination result with colleagues	1.80	Incentive	-
1	It makes a diagnosis for which I am not competent	1.71	Incentive	young (p=0.008)
2	It makes a diagnosis for which no specialist is accessible within an acceptable time or distance	1.59	Incentive	-
4	It saves me time	1.47	Incentive	young (p=0.004) men (p=0.02)
11	It allows me to manage the patient entirely in the office without handing over	1.40	Incentive	-
8	It allows me to predict a risk category of the patient	1.26	Incentive	-
6	It detects a specific pathology in an asymptomatic subject	1.17	Incentive	young (p=0.04)
14	It allows me to involve the patient in the therapeutic management	1.10	Incentive	-
32	It reassures the patient	1.10	Incentive	young (p=0.01)
22	It sometimes provides me with training	1.09	Incentive	young (p=0.0002)
15	It removes me from a situation of doubt	0.99	Incentive	young (p=0.003)
12	It allows me to adapt the treatment to a specific sub-category of patients (personalized medicine)	0.98	Incentive	young (p=0.01) urban (p=0.05)
33	It allows me to cut short abusive requests from the patient (e.g.: certificates, work stoppages, etc.)	0.87	Incentive	young (p=0.01) men (p=0.02)
5	It simplifies a procedure that I already know how to perform	0.84	Incentive	young (p=0.03) urban (p=0.02)
29	It increases the time spent with the patient	0.72	Incentive	-
7	It allows me to delegate a task to the secretary	0.71	Incentive	-
26	It is possible to capture the information (e.g. image) and then analyze it asynchronously	0.51	Neutral	young (p=0.04)
25	It shows me how to capture information (e.g. positioning a probe)	0.37	Neutral	-
13	It has a good reputation with patient associations	0.24	Neutral	-
31	It is very expensive but allows me to be profitable after a few months	0.21	Neutral	young (p=0.003)
19	It reflects a "high tech" image of my practice	0.09	Neutral	-
30	It reduces time spent with the patient	0.04	Neutral	-
28	It requires specific training from the manufacturer	-0.23	Neutral	-
18	It entertains the patient	-0.26	Neutral	-
24	It uses exclusively my personal smartphone	-0.36	Neutral	non-urban (p=0.03)
17	It gives information on elements for which it has not been requested (digression faculty)	-0.42	Neutral	-
27	It can be used freely in the waiting room by patients and their companions	-0.98	Deterrent	-
23	It automatically triggers an emergency call in some cases	-1.05	Deterrent	men (p=0.03)
16	It decides automatically without my intervention	-1.29	Deterrent	women (p=0.01)
3	It makes a diagnosis for which I am already competent	-1.39	Deterrent	-

"Deterrent items" (Table 2) relate to AI modules that: automatically trigger an emergency call, automatically decide without the GP intervention, make a diagnosis for which the GP is already competent, or can be used by the patient in the waiting room.

"Neutral items" (Table 2) relate to:

- Reputation with patients, "high tech" image of the practice, patient entertainment
- The way the information has to be captured, the need for training, the use of personal smartphones
- The economic model of the device
- The digression ability of the device

Table 2 also shows that many items are preferred by young GPS, and that some items are preferred according to the gender or the practice location.

Figure 1 enables to visualize the proportion of each answer (shortened wordings were used; refer to Table 2 to read complete wordings).

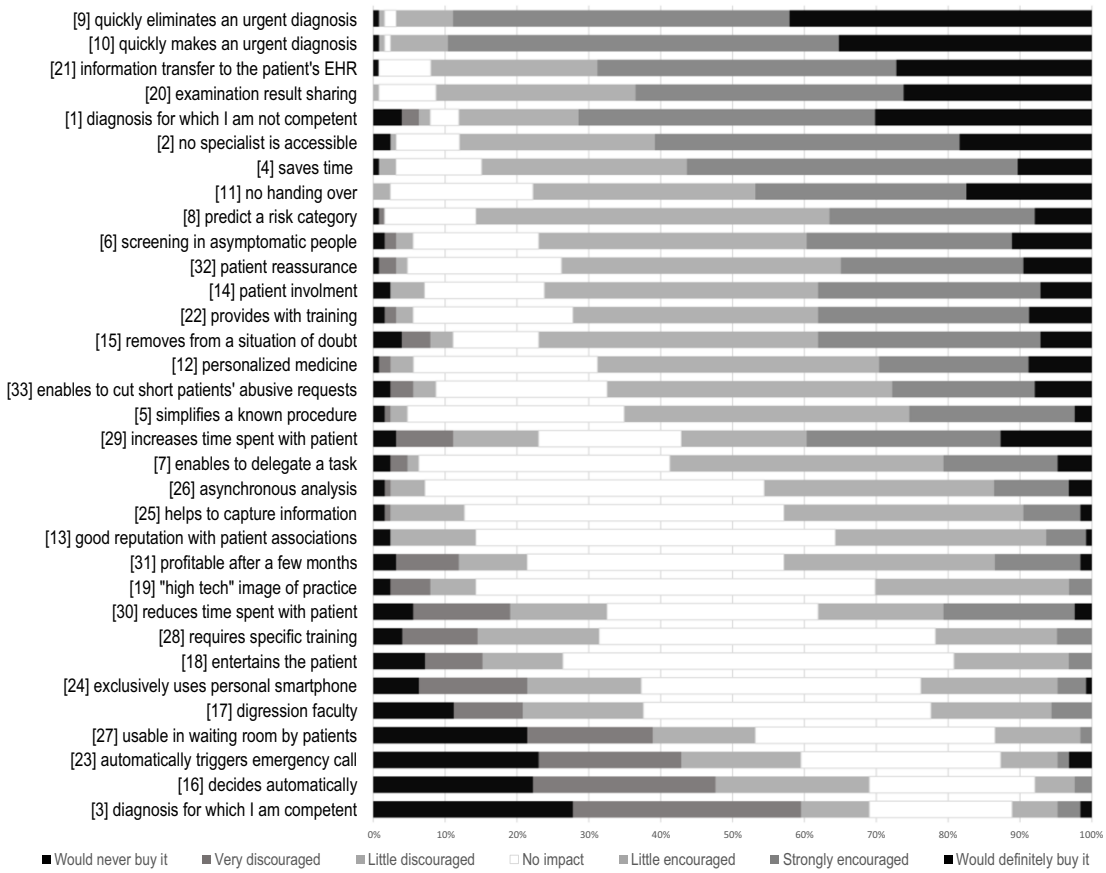

Figure 1 – Results of Likert scales: "Would this feature encourage you to, or discourage you from purchasing this module?". For original wording of items, refer to Table 2.

Discussion

We conducted a survey among GPs to find out what criteria would make them buy or refuse to buy a device with an AI module that could be used to support care.

As a consequence of the random draw and the high response rate, the characteristics of the surveyed population show a real representativeness of the French GP population.

Regarding the results, first criteria were in relation with the disease the module could enable to diagnostic: GPs expect such a module to diagnose or eliminate an urgent pathology for which they are not competent and for which specialists are difficult to access. They do not want confirmation of a diagnosis that they are already able to make. They are not specifically interested on digression ability.

There is also a strong demand from GPs for interoperability, with results integrated into the patient's EHR and easy transmission to colleagues.

GPs would like such modules to help them save time and simplify some procedures, or that enable them to delegate tasks to the secretary (but not to the patient him/herself), but they do not want such modules to automatically make a call or any action without confirmation.

As far as relations with the patient are concerned, GPs expect such a module to allow them to associate the patient with the care, to reassure him/her, but they do not pay particular attention to the image that the module sends back to the patient, and they do not want the module to be used by the patient alone in the waiting room. They are also looking for modules that allow them to personalize the care. Interestingly, GPs also want to be able to rely on the neutrality of a machine to cut off abusive requests, such as work stoppages or certificates of convenience.

Conclusions

We believe that, in addition to taking advantage of technical opportunities enabled by research progress, manufacturers should also prioritize AI module developments according to what end-users, especially general practitioners, expect from future AI modules.

Acknowledgements

This research did not receive any funding. Authors have no conflict of interest to declare.

References

[1] R. Colom, S. Karama, R.E. Jung, and R.J. Haier, Human intelligence and brain networks, *Dialogues Clin. Neurosci.* **12** (2010) 489–501.

[2] K. Benke, and G. Benke, Artificial Intelligence and Big Data in Public Health, *Int. J. Environ. Res. Public. Health*. **15** (2018). doi:10.3390/ijerph15122796.

[3] Journal of Medical Internet Research - Role of Artificial Intelligence Applications in Real-Life Clinical Practice: Systematic Review, (n.d.). https://www.jmir.org/2021/4/e25759/ (accessed May 7, 2021).

[4] C. Zheng, T.V. Johnson, A. Garg, and M.V. Boland, Artificial intelligence in glaucoma, *Curr. Opin. Ophthalmol.* **30** (2019) 97–103. doi:10.1097/ICU.0000000000000552.

[5] H. Shimizu, and K.I. Nakayama, Artificial intelligence in oncology, *Cancer Sci.* **111** (2020) 1452–1460. doi:https://doi.org/10.1111/cas.14377.

[6] T.M. Fonseka, V. Bhat, and S.H. Kennedy, The utility of artificial intelligence in suicide risk prediction and the management of suicidal behaviors, *Aust. N. Z. J. Psychiatry*. **53** (2019) 954–964. doi:10.1177/0004867419864428.

[7] R. Vaishya, M. Javaid, I.H. Khan, and A. Haleem, Artificial Intelligence (AI) applications for COVID-19 pandemic, *Diabetes Metab. Syndr.* **14** (2020) 337–339. doi:10.1016/j.dsx.2020.04.012.

[8] T. Grzywalski, M. Piecuch, M. Szajek, A. Bręborowicz, H. Hafke-Dys, J. Kociński, A. Pastusiak, and R. Belluzzo, Practical implementation of artificial intelligence algorithms in pulmonary auscultation examination, *Eur. J. Pediatr.* **178** (2019) 883–890. doi:10.1007/s00431-019-03363-2.

[9] P.I. Dorado-Díaz, J. Sampedro-Gómez, V. Vicente-Palacios, and P.L. Sánchez, Applications of Artificial Intelligence in Cardiology. The Future is Already Here, *Rev. Espanola Cardiol. Engl. Ed*. **72** (2019) 1065–1075. doi:10.1016/j.rec.2019.05.014.

[10] C. Delrot, G. Bouzillé, M. Calafiore, M. Rochoy, B. Legrand, G. Ficheur, and E. Chazard, Do Medical Practitioners Trust Automated Interpretation of Electrocardiograms?, *Stud. Health Technol. Inform.* **264** (2019) 536–540. doi:10.3233/SHTI190280.

[11] S. Keel, P.Y. Lee, J. Scheetz, Z. Li, M.A. Kotowicz, R.J. MacIsaac, and M. He, Feasibility and patient acceptability of a novel artificial intelligence-based screening model for diabetic retinopathy at endocrinology outpatient services: a pilot study, *Sci. Rep.* **8** (2018) 4330. doi:10.1038/s41598-018-22612-2.

Address for correspondence

Pr Emmanuel Chazard ; CERIM, faculté de Médecine, F-59045 Lille Cedex, France ;
emmanuel.chazard@univ-lille.fr ;
Phone: +33 3 20 62 69 69.