

A computational validation for the Health Concept Maturity Levels Questionnaire

Arthur Trognon

arthur.trognon@clinicog.fr

CLINICOG

Islem Habibi

CLINICOG

Hamza Altakroury

CLINICOG

Loann Mahdar-Recorbet

CLINICOG

Alizée Poli

CLINICOG

David Servais

Forum des Living Labs en Santé et en Autonomie

Julien Lambert

Association Innov'Autonomie

Mathias Béjean

Consortium DynSanté

Denis Abraham

Association Innov'Autonomie

Research Article

Keywords: Concept Maturity Level Health, Medical Technology, Computational Psychology, Computational Psychometrics, User-centered research, Psychological validation

Posted Date: July 10th, 2023

DOI: <https://doi.org/10.21203/rs.3.rs-3119800/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

The healthcare market is undergoing rapid transformation, requiring the integration of user needs from the earliest stages of product and service design. Living Labs are emerging as a model for the co-creation and evaluation of user-centered innovations. In this work, we developed a health CML grid and questionnaire to assess the maturity of health concepts.

Methods

The research process included multiple stages, starting with the creation of the Association Innov'Autonomie - Health Concept Maturity Levels Questionnaire – 178-items (AIA/CMLHQ178), designed to evaluate health concept maturity levels. Speech acts from Health CML expert interventions were then annotated and used as data for our machine learning and deep learning models. We used the CatBoost algorithm in the first experiment to discern individual Health CML factors from speech acts to generate factor probabilities used to feed a neural network trained to take the final decision, to evaluate whether the network could accurately identify the membership factors of Health CML criteria when presented with items from the AIA/CMLHQ178 questionnaire.

Results

The results of the study indicate that only the models trained with the true factors are able to correctly identify the corresponding factor in the sequentially encoded texts, with the exception of the need domains whose artificial performance was measured according to sensitivity. The general performance of the different CatBoost algorithms used to predict one factor versus the other two showed similar performance. For the questionnaire, the models trained with the real factors also showed better performance in identifying the matching factors compared to the random factors. A marginal difference was observed between the "Need" and "Technology" factors.

Conclusion

This study provides initial evidence of content validity for the AIA/CMLHQ178, introducing a novel approach to validate psychometric instruments using machine learning and deep learning techniques. However, overlaps between "Programmatic" and "Need" factors indicate a need for improvement in the CML Health model. Future work will focus on enhancing these models and investigating their potential application in other psychometric tools.

Background

The healthcare market is undergoing a rapid transformation, driven by technological advances and growing user expectations. Historically, technological innovation has been the main driving force behind the development of new products and services. However, it is now increasingly clear that the match between user needs and proposed solutions is essential to ensure the acceptance, adoption, and long-term success of these innovations¹. This market evolution highlights the need to integrate user needs from the earliest stages of the design and development of healthcare products and services.

Living Labs² have thus emerged as a promising model for addressing the importance of taking user needs into account in the development of healthcare products and services. This model is characterized by a user-centered, co-creation approach, where end-users, researchers, companies and decision-makers collaborate to identify, develop and evaluate innovative solutions³. This approach aims to ensure that innovations truly meet users' needs and are adapted to the contexts in which they will be deployed, thus contributing to wider adoption and dissemination. However, the ecosystem as a whole remains fractured, and a need for readability of the intrinsic properties of projects has emerged consensually among MedTech ecosystem players.

In response to these challenges, the CML (Concept Maturity Levels) Health grid⁴ has been developed to assess and guide the development of innovations in the healthcare sector. This grid takes into account technological, programmatic and, above all, user needs to assess the maturity and relevance of innovations. The main objective of this grid is to provide a framework for identifying gaps and areas for improvement, and to facilitate the decision-making process for players involved in the selection, development or commercialization of innovative healthcare solutions.

The CML Health questionnaire is an assessment tool based on the CML Health grid, designed to quickly and accurately measure the maturity of healthcare concepts. The grid examines three main areas: technological maturity, programmatic maturity, and consideration of user needs. Each of these axes is then broken down into sub-factors to provide a more in-depth and nuanced analysis of concept maturity. In the technological maturity axis, sub-factors include technology development, data management and intellectual property. For programmatic maturity, sub-factors include project management, regulatory and financial aspects. Finally, the axis of needs awareness is divided into sub-factors such as usage, market analysis and clinical evidence. The CML Health questionnaire is made up of several questions that address these axes and sub-factors, enabling players involved in the development and commercialization of innovative healthcare solutions to better understand the strengths and weaknesses of their project, and identify potential areas for improvement. The results of the questionnaire provide an overall view of concept maturity, facilitating informed decision-making and the prioritization of resources to maximize the impact and success of innovations in the healthcare sector.

In this article, we present an experiment to validate the Health Concept Maturity Levels Questionnaire, a tool based on the CML Health grid, which assesses the maturity of health concepts according to specific criteria and supports informed decision-making by project leaders and funders.

In general, the concept of psychometric validity encompasses three major aspects: construct, content and criterion validity. Construct validity determines the extent to which the proposed test can identify the construct being measured. It is generally obtained by correlating the measure with a number of other measures whose correlation patterns are theoretically predictable⁵. In contrast, to construct validity, content validity verifies the representativeness of the items in a given instrument in relation to the construct under study, as well as the degree to which this instrument possesses an appropriate sample of items for the measurement of the construct under analysis⁶. Finally, criterion validity aims to assess the performance predicted by the measurement tool. It is supported in particular by predictive validity, which seeks to verify the test's performance with regard to the criterion taken as the object of study, for example by predicting a diagnosis from the set of variables (i.e. items) considered⁷.

Recently published best-practice recommendations for questionnaire validation suggest that, in validation studies, a sample size of at least 10 participants per questionnaire item is expected^{8,9}. More generally, simulations have shown that at least 300–450 responses are expected in order to perform appropriate factor analyses¹⁰. In addition, it is recommended to use cross-sectional rather than longitudinal data^{9,11}.

In the present case, it didn't seem realistic to be able, at a given point in time, to collect enough data to carry out a validation by conventional methods: indeed, by superimposing available data from the literature, a sample-size of 450–1780 (10:1 ratio) respondents would have been necessary. However, only 287 companies would theoretically have been able to respond to the questionnaire at its highest maturity levels¹², thus limiting the interpretability of the results to only the weakest maturity measures. In addition, as the data was collected at the pace of MedTech events, we would not have been able to obtain cross-sectional data, thus limiting the generalizability of the results.

Therefore, in order to overcome these limitations, we decided to carry out a computational psychometrics experiment, an analytical framework inspired by the rationale of experimental design in animal research, and based on the comparative reading of algorithmic predictions to control the informational component of certain groups of variables that may represent latent factors or psychological constructs¹³. This framework has already been used to perform construct validity and criterion validity in quantitative data^{13,14}.

In the present work, we have adapted the usual computational psychometrics methodology to our limitations and aimed to build a computational content validation procedure using verbal material obtained in contexts of expert assessments and coaching sessions based on the CML Health model. These corpora were transcribed and sequenced, then annotated by assigning a criterion to each speech act produced by the interlocutors. These data were then used to train a hybrid pipeline composed of a layer of CatBoost algorithms and a layer of artificial neural networks to predict the respective attribution factors of the CML Health questionnaire items, thus making it possible to verify the extent to which these items are representative of the concepts they measure. In keeping with the canonical presentation of computational psychometrics, we constructed a negative control experiment to ensure that the performance measures we performed truly represented validity evidence and not over-fitting on noise. We thus created a second experiment in which we randomly and equiprobably assigned factors to items in the corpus, making it possible to verify the impossibility of predicting the questionnaire's attribution factors when the data are not encoded in accordance with the grid of the CML Health model.

METHODS

Model

The French CML Health (CMLH) model^{3,15} is a 9-levels reading grid that iteratively breaks down the innovation process into three interconnected axes: needs, technology, and programmatic. It is a direct descendant of the original CML model developed by the American National Aeronautics and Space Administration¹⁶, which incorporated the last two domains but now includes an additional user-centered axis. Furthermore, it specifies the technology and programmatic domains to adapt them to French and European regulatory requirements in research methodology and data management (Table 1)

Domain	Sub-domain
Needs	Usage
	Market
	Clinical proofs
Technology	Technical development
	Data management
	Intellectual property
Schedule	Project management
	Regulation
	Funding

Table 1 : Factorial structure of the French CML Health model, from (Trognon, 2023)

The first axis of the CML Health model (Table 2) is derived from the Technology-Readiness Level model¹⁷. It evaluates the development of technological concepts, product management, and ownership through three formalized axes: technological development, data management, and intellectual property. The technological development axis gradually assesses processes on a scale of 1 to 9, starting from evaluating the current state of the art through critical

functionality simulations, up to managing the product life cycle. The second axis focuses on how project leaders will handle data from their own devices, encompassing R&D data, protection protocols, and automation of product life cycle data. Lastly, the intellectual property axis provides insights into competitive analysis, monitoring existing patents, and managing potential infringements.

	CML1	CML2	CML3	CML4	CML5	CML6	CML7	CML8	CML9
Technological development	Evaluation of state-of-the-art	Conceptualization and theoretical analysis	Functional simulation and testing	Creation of software demonstrator	Development of alpha prototype	Technological analysis for improvement	Automation of function testing	Addressing software bugs and issues	Managing intellectual property
Data management	/N/A	Collection of R&D data	Organization of software data	Cybersecurity measures	Availability of data	Utilization of clinical data Intellectual property of clinical data	Access to data servers	Implementation of data collection devices	Managing intellectual property
Intellectual property	Monitoring of patents	Patents pending approval	Specific granted patents	Ensuring freedom of operation	Specific granted patents	Ensuring freedom of operation	N/A	N/A	Managing intellectual property

Table 2 : Example of the different maturity levels for each sub-area of the "Technological Maturity" domain, adapted from (Trognon, 2023)

The second axis of the CML Health model (Table 3) represents the first adaptation of the CML model developed by the American National Aeronautics and Space Administration. It incorporates project management and regulatory aspects, but in this specific case, tailored to European and French requirements for devices (such as access to the market under CE mark certification) and research. The aim is to adhere to Good Clinical Practices (GCP) and ethical principles safeguarding individuals participating in research, protecting them from potential risks associated with acquiring new biological or medical knowledge. It should be noted that research is conducted on healthy or sick volunteers with the intention of advancing knowledge in the biological or medical fields. The French regulatory framework is based on European regulations, with recent updates for innovation seeking medical device status (ANSM's proposed "clinical investigations" categories). Furthermore, it ensures that methods for collecting and processing health data comply with the General Data Protection Regulation (GDPR) and French reference research methodologies (MR-00X), which range from level 1 to 3.

	CML1	CML2	CML3	CML4	CML5	CML6	CML7	CML8	CML9
Project management	Identifying the driving factors	Conducting initial project risk analysis	Setting up test beds	Identifying complementary skills	Creating a detailed development plan	Updating project elements and addressing risks	Identifying marketing and sales skills	Finalizing and closing the project	Reviewing industrial development partners
Regulation	Establishing regulatory framework	Ensuring compliance with RGPD	Analyzing product risks	Assessing ethical aspects of the product	Collecting regulatory data	Consolidating the technical file	Compiling the CE mark file	Defining regulatory framework for data use	Renewing the CE marking
Funding	Identifying potential funding sources	Preparing the business plan	Planning financing for project demonstrators	Formalizing the business plan	Developing financial models	Establishing a Minimum Viable Business Model	Initiating Series A Capital Raising	Updating economic assumptions with real-life data	/

Table 3 : Example of the different maturity levels for each sub-area of the "Technological Maturity" domain, adapted from (Trognon, 2023)

The Programmatic axis of the CML Santé France model assesses the programmatic maturity of the project across three areas: project management, regulatory aspects, and financial aspects (Table 4). The project management axis evaluates the project consortium, from pilot identification to development partnership renewal, including the creation of Test Beds, and examines the nature of partnerships formed. The regulatory axis assesses programmatic maturity from the initial legal investigation surrounding the project to CE mark renewal, encompassing product risk analysis, compliance with European (e.g., MDR, GDPR) and French (i.e., ethics, clinical investigations for medical devices) regulatory constraints. Lastly, the financing axis allows for gradual evaluation of financial aspects, ranging from identifying potential financing sources to updating business economic assumptions based on real-life usage data of the device.

	CML1	CML2	CML3	CML4	CML5	CML6	CML7	CML8	CML9
Usage	Understanding social and public health context	Identifying practice situations that justify the need	Collaboratively developing tailored usage scenarios	Conducting UX/UI lab evaluations	Defining the usage industrialization scheme	Assessing usability and acceptability	Evaluating ecological impact of a pre-series	Examining real-life organizational impact	Ensuring patient-reported experience
Market	Reviewing the existing market literature	Identifying the value proposition	Defining product positioning and expected impact	Quantifying the expected impact	Developing market access strategy	Characterizing the device based on usage surveys	Implementing marketing elements (deployment, export)	Refining go-to-market strategies by customer type	Market access differentiation
Clinical	Reviewing relevant clinical literature	Identifying the medical need	Formulating the clinical strategy	Initiating preliminary clinical trials	Analyzing results from clinical trials	Drafting study reports (publications)	Conducting multi-center clinical trials	Performing medico-economic studies	Ensuring patient-reported outcomes

Table 4 : Example of the different maturity levels for each sub-area of the "Need Maturity" domain (PREM : Patient Reported Experience Measure ; PROM : Patient Reported Outcomes Measure), adapted from (Trognon, 2023)

Finally, the last axis of the CML Santé France model represents the true innovation of the consortium by specifying the CML model as described by NASA. It incorporates elements of consumer behavior theory¹⁸, particularly addressing the barriers to innovation, and enables the evaluation of maturity in terms of needs across three axes: uses, market, and clinical evidence (Table 5). The Uses axis provides insights into the device's value and ensures user-centric product development in terms of uses. It assesses development from identifying the social context and public health implications to evaluating the perceived quality of care through patient evaluation methods (PREM). This axis verifies the elimination of certain functional barriers, such as conflicts with established usage patterns. The Market axis examines the competitive landscape concerning device uses, including market literature reviews and evaluations of market segment diversity and respective access strategies. This axis confirms the elimination of functional value barriers and verifies the uniqueness of the device's value proposition. Finally, the Clinical Proof axis assesses the quality of clinical investigations conducted on the device, ranging from comprehensive literature analysis to formalized processes for evaluating the perceived quality of device results by patients (PROM). This axis addresses the functional barrier of uncertainty that arises when end-users have limited access to devices under development, as described by Stone and Grønhaug¹⁹.

Subjects

Participant data were collected during sessions organized between 2021 and 2023, including auditions of innovative project leaders in the field of medical technologies (MedTechs). In chronological order, the data comes from two individual semi-directive interviews carried out as part of the preparation of the methodological deployment; five consortium auditions organized during the Future4Care 2022 start-up competition; three semi-directive group interviews carried out during the French National Digital Health Sector's Call for Expressions of Interest; and a day of presentations (symposium) as part of the e-Meuse Expert Committee meeting, whose aim is to identify the optimal conditions for the deployment of digital innovation in healthcare in the Eastern France.

All participants received detailed information on the objectives and purpose of the study and ethical consents were obtained online in agreement with the Declaration of Helsinki and the study protocol was approved by the Institutional Review Board Commission Nationale de l'Informatique et des Libertés (registration n°2230503).

Study design

The verbatim of pitches, individual and group interviews were recorded, then transcribed according to the formalism of the 2TK model^{20,21}, i.e. transcribed into speech acts sequenced by the pauses identified in the discourse in order to reproduce its kinetics as closely as possible; as well as to identify the most circumscribed packets of information possible. Each speech act was then labeled according to 3 parameters: the factor (three levels: need, programmatic, technology; otherwise "Null"); the sub-factor (three levels per factor: market, uses, clinical evaluation for the "Need" factor; technological development, intellectual property, data management for the "Technology" factor; and financing, regulatory aspects, project management for the "Programmatic" factor); and the CML level (nine levels: CML1 to CML9). n=10952 speech acts were labeled; then only non-Null speech acts (n=2070) were retained.

These speech acts were used to build a two-layer computational pipeline with two complementary and distinct objectives: a first "expert systems" layer of CatBoost^{22–24} algorithms, a per-factor algorithm for calculating the probability of the speech act carrying information about the given CML Health factor or belonging to another factor or noise; and a second layer consisting of an artificial neural network embedding the probability of each expert system and providing a final decision regardless of the textual material provided, and which will be used to predict the membership factor of the questionnaire items from which the speech acts have been labeled.

Scale development & item generation

The items were developed directly from the Concept Maturity Levels Santé grid of the Forum Living Lab Santé Autonomie (version 2021/12; Supplementary File 1). We generated 1 item for each criterion, with the aim of eliciting a personal representation of the project's status, from which the participant could make a judgment between the induced representation and the subjective perception of progress. A total of 178 items were generated from this grid (French : Additional File 2 ; English : Additional File 3).

We chose to base the questionnaire on an Osgood-type scale, constructing the lower and upper bounds in a logic of continuity, where items belonging to higher levels have lower bounds that can be superimposed to some extent on the upper bounds of lower items. Similarly, we chose to scale the questionnaire on a linear 6-point scale in order to avoid neutrality bias, although its existence is still debated^{25,26}. In addition, the stimuli were built around first-person personal pronouns, which allows for greater immersion of the subject in relation to the situational context about which he or she is being questioned²⁷, which seems particularly appropriate in the case of a maturity assessment of project concepts.

However, not all sentences describing criterion within the original grid were unambiguous in terms of their specific factors, necessitating few tuning editions to ensure their alignment with the intended factor. For instance, the phrase “The technical tools that can meet the *needs* are clearly identified” has the factor “Technology” in the questionnaire, even though it contains the word “need”. In this case the criterion was modified to “The appropriate tools to meet the technical requirements are clearly identified.” in the scale. Once the item adaptation was completed, the corpus underwent annotation and classification was performed using the computational pipeline described in the relevant section.

Computational content validation

Computational content validation is an innovative procedure that uses computational techniques based on machine learning to assess the content validity of a measurement instrument, scale or questionnaire. It also aims to verify whether the instrument's items or questions adequately reflect the concept they are designed to measure.

Within the framework of computational content validation, an algorithmic model is trained on a portion of *verbatim* data obtained from discussions focused on project maturity, and its performance is then evaluated on a different portion of the data (called a test set). The aim is firstly to demonstrate that the model can reliably and accurately predict the dimensions or factors underlying the discourse and superimposable on the CML Health questionnaire criteria from the words spoken by the participant on a given sample; then to demonstrate that this model is capable of correctly identifying these dimensions in a new data set (i.e. the questionnaire), providing an additional segment of evidence for the questionnaire's content validity. This approach thus offers a new computational alternative to traditional content validation methods, which are usually based on a pre-test of the questionnaire, whereas in the case of computational content validity, this step can be carried out under pre-test conditions.

Factorial structure and content validation

The method employed in this study aims to circumvent the limitations inherent in classical psychometric validation by using a computational approach. Here's how it could help overcome these constraints while providing relevant information on factor structure, internal consistency and reliability.

Factorial structure: Using a machine learning model, we are able to identify which factors (in our case, the categories of 'Technology', 'Programming' and 'Need') are relevant to each statement. This is an indirect way of exploring the factor structure of the dataset. Indeed, if the model is able to predict these factors correctly, it suggests that they are well-defined and have real meaning in the context of technological innovation in healthcare.

Internal consistency and reliability: Internal consistency and reliability are assessed through the performance of the machine-learning model. In this context, internal consistency is reflected by the model's ability to make consistent predictions on different parts of the training dataset. Reliability, on the other hand, is assessed by examining the model's performance on a novel test dataset. If the model makes accurate and consistent predictions on this data set, this suggests that it is reliable.

This approach has the advantage of not requiring a large number of participants, which is a major constraint of classical psychometric validation. What's more, the machine learning model can be re-trained as new data are collected, enabling continuous improvement in accuracy and reliability.

It's important to note that this computational approach is not a substitute for conventional psychometric validation. On the contrary, it offers a practical alternative when the data collection required for conventional validation is not feasible. In addition, it allows conclusions to be drawn from a fundamentally smaller data set, but one that is potentially richer in information, offering the possibility of continuous updating as new data are collected. It can also be superimposed on conventional validation techniques, providing a new body of evidence for the validity of the measurements produced by these tools.

Computational content validation procedure

The data used in this study consisted of transcribed text derived from meetings and presentations inscribed by 2TK method^{20,21}. For performing the analysis, three experts assigned labels to the transcribed phrases based on the relevant factors from the CML Health model (Need, Programmatic and Technology). For representing the transcribed phrases in a format suitable for analysis, the Bag of Words technique was applied. This approach transformed the text into numerical vectors, enabling subsequent computational processing and modeling.

The proposed model in this work was to have two layers of classification (as shown in Figure 1): A layer which provides a probability that transcribed phrases belong to a certain factor in the CML metric ; while the aim of the second layer is to provide the predicted label of the transcribed phase based on the given probabilities and the corresponding transcribed phrase (figure showing the plan in our study).

Figure 1: Computational analysis pipeline

In order to determine the most suitable classifier for our study, a comparison was performed among various classifiers, including Decision Tree, Bagging classifier (comprising 500 decision trees), Extra Tree, Gradient Boosting, AdaBoost, Catboost, and Random Forest (Figure 2). The comparison was conducted using a 2/3 train and 1/3 test split of the data. This evaluation allowed us to identify the classifier that exhibited the highest performance and accuracy. Following the evaluation, the Catboost classifier emerged as one of the top-performing classifiers, demonstrating a high level of accuracy in classifying the

transcribed phrases. The decision to select Catboost was further supported by its flexibility in handling categorical data, which was a relevant characteristic for our study.

Figure 2: Model selection for building the computational pipeline.

To optimize the performance of the Catboost classifier, the hyperparameters were fine-tuned using Bayesian Optimization. This optimization approach ensured that the hyperparameters of the classifier were optimized to obtain improved classification results. These hyperparameters (and their corresponding ranges) were: The number of estimators ([100, 1500]), the learning rate ([0.01, 0.3]), the depth of each decision tree ([4, 10]), the L2 regularization coefficient ([1, 10]), the randomness strength ([0, 1]), the temperature parameter for the bagging process ([0, 10]) and the border count parameter for categorical features ([32, 255]). Furthermore, the number of initial points for the Bayesian Optimization was set to five and the number of iterations was set to 20. Subsequently, the Catboost classifier was utilized to generate probabilities for each category (Pr(Need), Pr(Programmatic) and Pr(Technology)) associated with the transcribed text. These probabilities served as an indication of the likelihood of each category's presence in the transcribed phrases.

Furthermore, in order to enhance the classification process, a Neural Network was trained using the probabilities generated by the Catboost classifier, in addition to the transcribed text. Another set of data was reserved for testing the performance of this Neural Network layer.

RESULTS

CatBoost "Expert" Algorithms Train-Testing:

Results of the classification performances between real CML Health factor labels and random ones can be observed in Figure 3. Data suggest that only models trained using the real factors are able to correctly identify the matching factor on the basis of sequentially encoded text corpora, except for needs domains whose artifactual performance has been measured on the sensitivity metric.

Figure 3. The average performance of each factor expert CatBoost in identifying the membership factor when a speech act related to the CML Health model is presented for each condition in 5-Fold cross-validation.

Regarding the general accuracy measures, a two-factor ANOVA (Performance~Condition*Algorithm) the different algorithms used to predict one factor in contrast to the two others showed similar performances [$F(2,24) = 0.606, p = 0.554$]; and the randomization of the factor for the control condition resulted in a critical performances loss [$F(1,24)=1122.131, p<.001$]. A significant interaction was detected between the different CatBoosts performances and the condition [$F(2,24) = 62.604, p < 0.001$].

Further post-hoc analysis revealed a significant difference between experimental and control conditions for all CatBoost algorithms (Table 5, Figure 3 (Left).

CatBoost algorithm	Experimental VS Control (<i>p</i> -value)
"Need" identifier	<.001
"Programmatic" identifier	<.001
"Technology" identifier	<.001

Table 5 : Tukey post-hoc test results for the accuracy (algorithm*condition interaction)

Concerning the detection measures, a three-factor ANOVA (Performance~Condition*Algorithm*Metric) revealed a significant effect of metric [$F(1,48)=441.298, p<.001$] and group [$F(1,48)=412.299, p<.001$]. Significant interactions were observed between the algorithms and the condition [$F(2,48)=10.435, p<.001$], metric and condition [$F(2,48)=10.290, p=.002$], and between condition, algorithm and the metric considered [$F(2,48)=122.204, p<.001$]. Further Tukey's post-hoc analysis revealed significant differences between experimental and control conditions for all CatBoost Algorithms, except for the sensitivity in the need domain and for the specificity in the programmatic domain (Table 6, Fig. 3).

CatBoost algorithm	Sensitivity (Exp vs Ctrl)	Specificity (Exp vs Ctrl)
"Need" identifier	.45	<.001
"Programmatic" identifier	<.001	.96
"Technology" identifier	<.001	<.001

Table 6 : Tukey post-hoc test results for the metrics (algorithm*condition*metric interaction)

These results can be explained by the overlap between the "Regulatory" sub-factor of the "Programmatic" factor and the "Clinical Evaluation" sub-factor of the "Need" factor, both of which share criteria concerning ethical questioning and subsequent evaluation by human protection committees in order to initiate clinical investigations.

Neural Network Performance on the Questionnaire:

Results of the classification performances for the questionnaire factor identification between the conditions implying real CML Health factor labels and random ones can be observed in Fig. 4. Data suggest that only models trained using the real factors are able to correctly identify the matching factor when an item of the questionnaire was presented.

Figure 4. The average performance of the neural network in identifying the membership factor of each CML Health questionnaire item for each condition in 5-Fold cross-validation.

Regarding the general accuracy measures, a two-way ANOVA (Performance~Condition*Factor) revealed that the neural network showed almost similar performance for all factors [$F(2,24)=3.24, p=.006$], except between the factors "Need" and "Technology" where a marginal difference was measured using the same post-hoc Tukey's procedure [$p=.05$]. As previously, a strong difference was observed in terms of performances between the experimental and control conditions [$F(1,24)=123.52, p<.001$], but no interaction between the variables under consideration has been measured.

Concerning the detection measures, the three-way ANOVA (Performance~Condition*Factor*Metric) revealed a significant effect for the metric [$F(1,48)=130.81, p<.001$] and the experimental condition [$F(1,48)=147.593, p<.001$], with significant interactions between the target factor and the metric used to assess the performance [$F(2,48)=5.872, p=.005$], the experimental condition and the metric used to assess the performance [$F(1,48)=17.591, p<.001$], and between the three variables under consideration [$F(2,48)=14.6, p<.001$]. Further, post-hoc Tukey's analysis confirmed as intended a significant difference between experimental and control conditions for both sensitivity [$p<.001$] and specificity [$p<.001$].

Concerning (Target Factor x Experimental Condition) interaction specifically, a subsequent post-hoc Tukey procedure revealed significant differences between sensitivity and specificity for all three target factors, suggesting that its abilities to identify and exclude different target factors are not equivalent (Table 7)

Target factor	Sensitivity VS Specificity (<i>p</i> -value)
Need	<i><.001</i>
Programmatic	<i><.001</i>
Technology	<i><.001</i>

Table 7 : Tukey post-hoc test results for the (Target Factor x Experimental Condition) interaction

Finally, the Tukey post-hoc analysis investigating the triple (Target Factor x Experimental Condition x Metric used) interaction showed an opposite pattern in regard to the first experience (Table 8).

Target factor	Sensitivity (Exp vs Ctrl)	Specificity (Exp vs Ctrl)
Need	<i><.001</i>	.99
Programmatic	.11	<i>.005</i>
Technology	<i><.001</i>	<i><.001</i>

Table 8 : Tukey post-hoc test results for the (Target Factor x Experimental Condition x Metric used) interaction

These results again suggest overlaps between the need, and programmatic, factors. The pattern inversion between the two experiments is an interesting observation that may reinforce the idea of a strong overlap between the two factors.

DISCUSSION

Using a hybrid approach involving supervised machine-learning and unsupervised deep-learning approaches, we were able to provide evidence of content validity for the Association Innov/Autonomie - Health Concept Maturity Levels Questionnaire – 178-items (AIA/CMLHQ178).

First, we carried out an experiment using a series of CatBoosts, state-of-the-art supervised classification algorithms that verified that it was possible to identify the Health CML factor specifically emerging from the noise when a speech act was presented.

These probabilities were then used to run an experiment using an artificial neural network, which showed that it was possible to identify the membership factors of Health CML criteria when presented with questionnaire items.

To the best of our knowledge, and in spite of the evidence provided for the validity of the questionnaire itself, we believe that this work is the first to investigate the validity of the Concepts Maturity Levels model itself, even though this model has already been built in several iterations before its French declension in the healthcare sector (ref)

What's more, these works are the first to our knowledge to investigate the principle of psychometric content validity computationally and using strictly qualitative data, thus providing a new layer of producible evidence within the overall framework of psychometric validity.

However, the results suggest room for improvement: indeed, the specificity and sensitivity measures for the "Programmatic" and "Need" factors were not satisfactory in the inversion experiments, suggesting overlaps between these two factors. This can probably be explained by the items measuring the ethical dimensions of the project, both shared by the "Clinical Evaluation" sub-factor of the "Need" factor in the criteria evaluating clinical evidence, notably in the constitution of clinical investigation files and the analysis of clinical study data in general; and both shared by the "Regulatory" sub-factor of the "Programmatic" factor, which evaluates the ethical framework associated with clinical investigations, notably the constitution of personal protection files, as

well as the planning of the clinical studies themselves. A re-allocation or modification of the initial grid thus seems necessary in order to increase the internal consistency of the Health CML model itself.

Several limitations of this study need to be acknowledged. First, the dataset used was relatively small, and the number of individual participants was limited. This restricted the opportunity for A-B testing on a fresh dataset due to the multiple splitting operations that were required in the machine learning processes. As such, our findings should be considered preliminary, and caution is required in generalizing them to other contexts.

Second, the study did not include a traditional quantitative measure on the questionnaire itself. While the hybrid approach of machine learning and deep learning provides a novel perspective on content validity, it does not replace canonical psychometric methodologies. The absence of a quantitative validation in our study limits the triangulation of evidence to support the questionnaire's validity.

Lastly, the study did not conduct a detailed analysis of the decision shifts that are to be considered in future studies. Understanding these shifts could potentially provide insights into the underlying mechanisms of how the algorithms identify the membership factors, providing information about the factorial structure as well. This may have resulted in an incomplete view of the behavior and performance of the models used. This lack of detailed analysis of decision shifts represents an important avenue for future research on Concept Maturity Levels and psychometric tools. Future studies should aim to address these limitations to provide a more comprehensive understanding of the performance of machine and deep learning methods in validating psychometric instruments.

CONCLUSION

Through a combination of supervised and unsupervised machine learning and deep learning approaches, we were able to provide the first evidence of the validity of the Health CML model, as well as of the content validity of the AIA/CMLHQ178 questionnaire. In addition, the method used provides a new tool for generating evidence of psychometric validity through a computational approach using qualitative data.

Declarations

Ethics approval and consent to participate

Consent was obtained online before the completion of the survey using four questions in the questionnaire.

- I am voluntarily participating in this research
- I can stop my participation at any time without having to give explanations
- All information I provide will be kept confidential and my identity will never be disclosed.
- I authorize the conservation and use of this confidential data for scientific research in psychology.

The study did not collect any identifiable data (no name, date of birth, location, or IP address), and only collected the time stamp of the pass (format: DD/MM/YY H:M:S). All participants received detailed information regarding the purpose of the study and provided online informed consent to participate in the study. All procedures were conducted in accordance with the principles of the Declaration of Helsinki and the study protocol was approved by the Institutional Review Board Commission Nationale de l'Informatique et des Libertés (CNIL - registration n°2230503).

Consent for publication

Not applicable

Availability of data and materials

The dataset used in the study cannot be shared due to confidentiality agreements signed with the companies that participated in the CML Health workshops and sessions organized by the Consortium. The full AIA/CMLHQ178 questionnaire is available in Additional File 2 (French) and 3 (English).

Competing interests

The authors declare that they have no competing interests

Funding

This study was funded by "Agence Nationale de la Recherche" (ANR-20-CE26-0015)

Authors' contributions

Contributor Role	Contributor
Conceptualization	Arthur Trognon, David Servais, Mathias Béjean
Data curation	Arthur Trognon, Alizée Poli, Islem Habibi, Hamza Altakrouy
Formal Analysis	Arthur Trognon, Hamza Altakrouy
Funding Acquisition	Mathias Béjean
Investigation	Arthur Trognon
Methodology	Arthur Trognon
Project Administration	Arthur Trognon, David Servais, Denis Abraham, Mathias Béjean
Resources	David Servais, Denis Abraham & Mathias Béjean
Software	Arthur Trognon, Hamza Altakrouy, Loann Mahdar-Recorbet
Validation	Arthur Trognon
Visualization	Arthur Trognon & Islem Habibi & Hamza Altakrouy
Writing - Original Draft Preparation	Arthur Trognon, Hamza Altakrouy & Islem Habibi
Writing - Review & Editing	Arthur Trognon & Islem Habibi

Acknowledgements

Not applicable

References

1. Laukkanen T. Internet vs mobile banking: comparing customer value perceptions. *Bus process Manage J.* 2007;13:788–97.
2. Leminen S, Westerlund M, Nyström. A.-G. Living labs as open-innovation networks. (2012).
3. Béjean M, Picard R, Bréda G. Living Labs, innovation collaborative et écosystèmes: le cas de l'initiative «Concept Maturity Levels» dans les Medtech. *Innovations.* 2021;65:81–110.
4. Béjean M, Siqueira E. Organizing Medtech Innovation with Concept Maturity Levels. in *HEALTHINF* 621–626 (2019).
5. Westen D, Rosenthal R. Quantifying construct validity: two simple measures. *J Pers Soc Psychol.* 2003;84:608–18.
6. Shi J, Mo X, Sun Z. [Content validity index in scale development]. *Zhong Nan Da Xue Xue Bao Yi Xue Ban.* 2012;37:152–5.
7. Taherdoost H. Validity and reliability of the research instrument; how to test the validation of a questionnaire/survey in a research. *How to test the validation of a questionnaire/survey in a research (August 10, 2016)* (2016).
8. Nunnally JC. *Psychometric theory.* (1967).
9. Boateng GO, Neilands TB, Frongillo EA, Melgar-Quinonez HR, Young SL. Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. *Front Public Health.* 2018;6:149.
10. Guadagnoli E, Velicer WF. Relation of sample size to the stability of component patterns. *Psychol Bull.* 1988;103:265–75.
11. Chesney MA, Neilands TB, Chambers DB, Taylor JM, Folkman S. A validity and reliability study of the coping self-efficacy scale. *Br J Health Psychol.* 2006;11:421–37.
12. Catégories d'entreprises – Tableaux de l'économie française | Insee. <https://www.insee.fr/fr/statistiques/3303564?sommaire=3353488>.
13. Trognon A, Cherifi YI, Habibi I, Demange L, Prudent C. Using machine-learning strategies to solve psychometric problems. *Sci Rep.* 2022;12:18922.
14. Trognon A, Richard M. Questionnaire-based computational screening of adult ADHD. *BMC Psychiatry.* 2022;22:401.
15. Trognon A et al. Shaping User-Centered Health Innovation Through Assessment. (2023).
16. Wessen R et al. Space mission concept development using concept maturity levels. in *AIAA Space 2013 Conference and Exposition* 5454 (2013).
17. Mankins JC. Technology readiness levels. *White Paper* April. 1995;6:1995.
18. Claudy MC, Garcia R, O'Driscoll A. Consumer resistance to innovation—a behavioral reasoning perspective. *J of the Acad Mark Sci.* 2015;43:528–44.
19. Stone RN, Grønhaug K. Perceived risk: Further considerations for the marketing discipline. *Eur J Mark.* 1993;27:39–50.
20. Trognon A. Computational diagnosis of Shwachman-Diamond syndrome through cognitive and dialogical investigations. *Lorraine University;* 2022.
21. Trognon A, Humeau C, Verhaegen F, Musiol M. Vectorial topological deviation as a measure of thought organization: insight from schizophrenic corpuses in free dialogue situations analyzed with the 2TK model of interaction. (2021).
22. Dorogush AV, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363* (2018).
23. Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. *J big data.* 2020;7:1–45.
24. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems* 31, (2018).

25. Lozano LM, García-Cueto E, Muñiz J. Effect of the number of response categories on the reliability and validity of rating scales. *Methodology: Eur J Res Methods Behav Social Sci.* 2008;4:73–9.
26. Weijters B, Cabooter E, Schillewaert N. The effect of rating scale format on response styles: The number of response categories and response category labels. *Int J Res Mark.* 2010;27:236–47.
27. Hartung F, Burke M, Hagoort P, Willems RM. Taking Perspective: Personal Pronouns Affect Experiential Aspects of Literary Reading. *PLoS ONE.* 2016;11:e0154732.

Figures

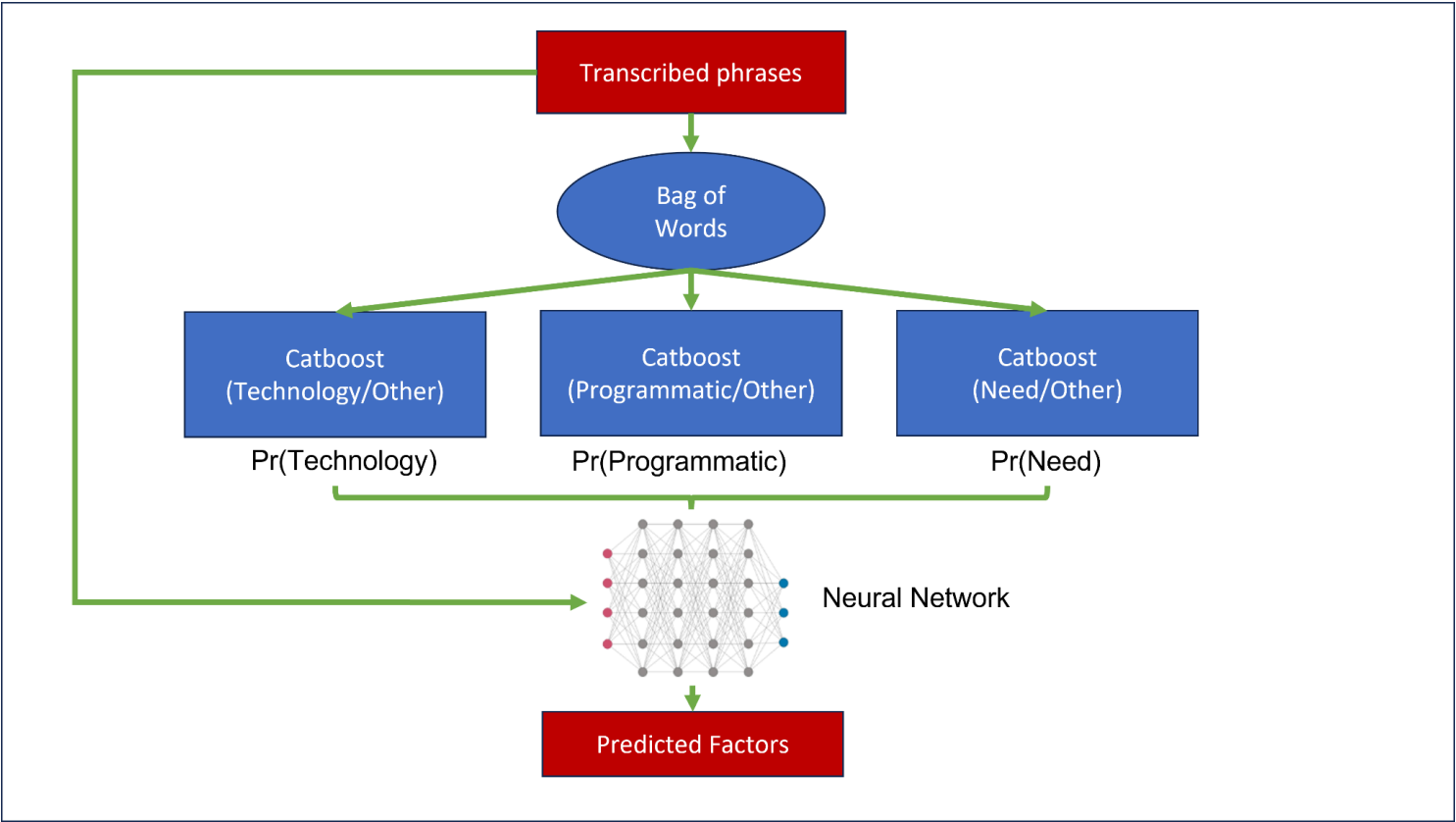


Figure 1

Computational analysis pipeline

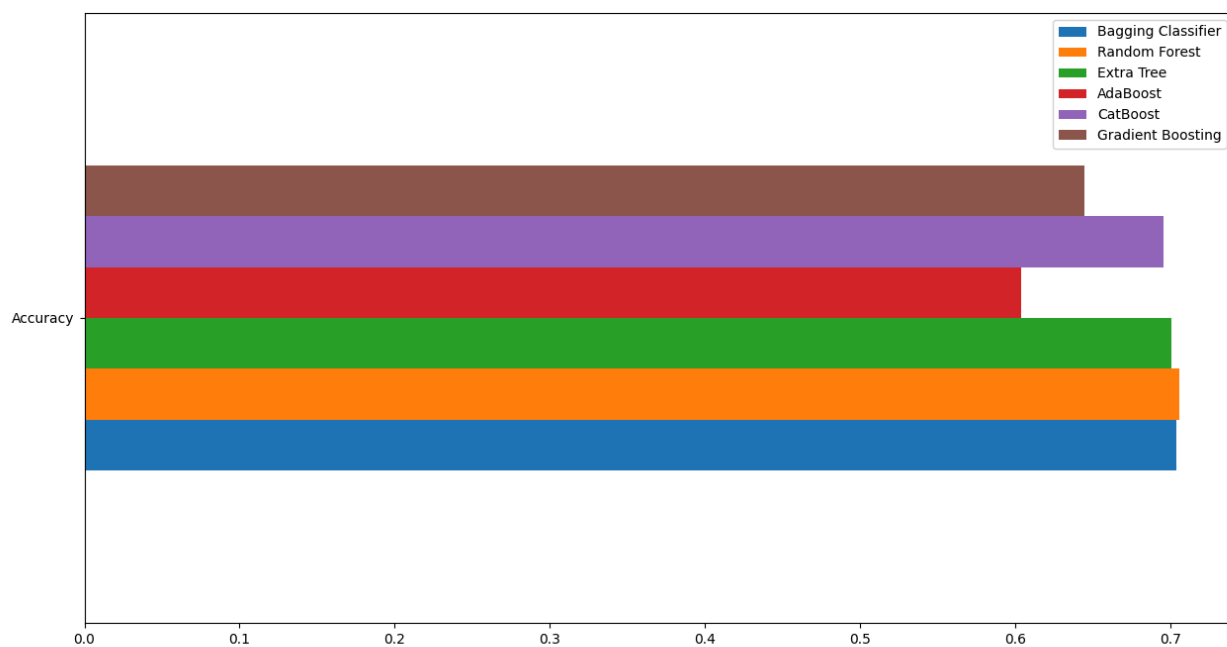


Figure 2

Model selection for building the computational pipeline.

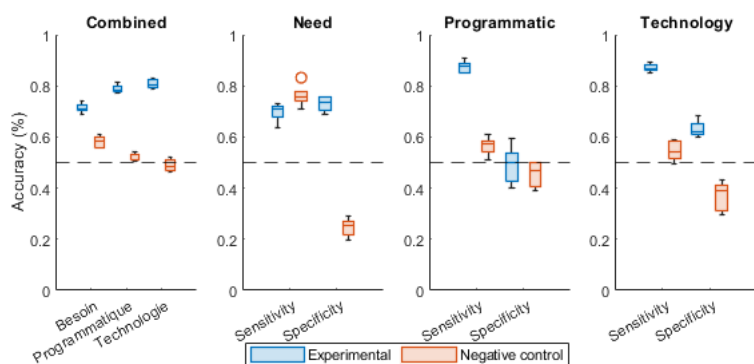


Figure 3

The average performance of each factor expert CatBoost in identifying the membership factor when a speech act related to the CML Health model is presented for each condition in 5-Fold cross-validation.

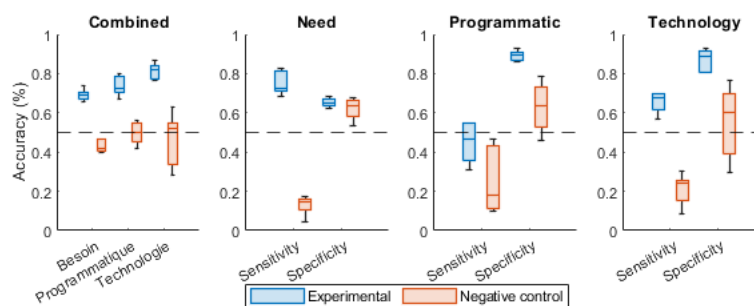


Figure 4

The average performance of the neural network in identifying the membership factor of each CML Health questionnaire item for each condition in 5-Fold cross-validation.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [cml1additionalFile1IlsaGrid.pdf](#)
- [cml1additionalFile2fullQuestionnairefrench.xlsx](#)
- [cml1additionalFile3fullQuestionnaireEnglishVersion.xlsx](#)