Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Public acceptance of artificial intelligence in health care: a cumulative hierarchy in a 16-item instrument, German–Italian measurement equivalence, and a five-item short form

Dietmar Ausserhofer
Claudiana — College of Health Professions
Christian J. Wiedermann
christian.wiedermann@am-mg.claudiana.bz.it

Claudiana — College of Health Professions
Verena Barbieri
Claudiana — College of Health Professions
Nicoletta Colletti
Provincial Institute of Statistics (ASTAT), Autonomous Province of Bolzano-Bozen
Stefano Lombardo
Provincial Institute of Statistics (ASTAT), Autonomous Province of Bolzano-Bozen
Timon Gärtner
Provincial Institute of Statistics (ASTAT), Autonomous Province of Bolzano-Bozen
Giuliano Piccoliori
Claudiana — College of Health Professions
Doris Hager von Strobele-Prainsack
Claudiana — College of Health Professions
Wolfgang Wiedermann
University of Missouri

Research Article

Keywords: artificial intelligence, health care, acceptance, Mokken scale analysis, nonparametric item response theory, differential item functioning, measurement equivalence, short form, latent class analysis, South Tyrol

Posted Date: September 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10966355/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.

Surveys of public attitudes towards artificial intelligence (AI) in health care report acceptance application by application, which cannot show whether acceptance forms an ordered cumulative structure, holds across language groups, or can be captured by few items.

Methods.

Data came from a probabilistic population panel in South Tyrol, Italy, surveyed February–March 2026 (N = 901). Respondents rated the acceptability of 16 AI applications; uncertainty responses were scored as non-acceptance. Mokken scale analysis assessed scalability, monotonicity, invariant item ordering and local independence. Differential item functioning (DIF) was assessed by logistic regression on the sum score, with the increment in McFadden pseudo-R^2 and false discovery rate adjustment; language group was the primary grouping. A short form was derived by exhaustive evaluation of all five- and six-item subsets; the selection rule was cross-validated over 200 random splits.

Results.

The applications formed a single moderately strong scale (H = 0.491), with monotonicity throughout. Invariant item ordering across respondents was not supported (H^T = 0.267): the scale orders groups reliably, individuals only in part. Item difficulty ordering was practically identical across subgroups (Spearman rho ≥ 0.962), and no item showed relevant DIF between the German and the Italian version (largest pseudo-R^2 increment 0.0070). A five-item short form achieved H = 0.597, better person ordering (H^T = 0.507, 97.7th percentile) and a correlation of 0.903 with the 16-item score. Its composition was not uniquely determined: the rule returned 34 forms across 400 selection–evaluation pairs, the adopted one once, and appointment scheduling, selected in 77.2%, is not among its items. Latent class analysis identified four acceptance types, one (29.7%) separating administrative from clinically consequential applications categorically rather than by degree; the class number lacked formal support.

Conclusions.

Acceptance of AI in health care in the general population follows a cumulative hierarchy ordered by the clinical consequence of the application. It is invariant across the German- and Italian-speaking population despite culturally adapted translations, supporting parallel-version administration. A five-item short form reproduces it, although which five items are selected is not uniquely determined. Whether the hierarchy and its invariance hold among health professionals requires further research.

Background

Artificial intelligence has moved from research settings into routine health care across diagnostic, therapeutic, administrative and communicative domains, and the pace of that movement has raised the

question of how the populations served by these systems evaluate them [1, 2]. A scoping review of the Italian evidence base identified two gaps of direct relevance to that question: the absence of probabilistic, population-representative surveys, and the absence of multilingual comparisons within the Italian regions in which more than one language community is served by the same health system [3].

Studies that measure public acceptance typically present the proportion of respondents endorsing each application, and such descriptive reporting has established a robust regularity: administrative applications attract more support than applications that bear on clinical decisions, and applications involving time-critical or high-stakes decisions attract the least [4–10]. What that reporting format cannot establish is whether the applications stand in a systematic relation to one another. Two survey results that are identical in their marginal proportions can arise from very different response structures. In one, respondents who endorse a demanding application also endorse all less demanding ones, so that acceptance is cumulative and a single ordering of applications describes the population. In the other, respondents endorse idiosyncratic subsets, so that the marginal proportions summarise unrelated evaluations. The distinction matters for measurement, because only the first case licenses a sum score as an interpretable quantity, and it matters substantively, because only the first case supports statements of the form that endorsement of one application implies endorsement of another [11]. General attitude scales for artificial intelligence are themselves explicitly multidimensional, separating positive from negative evaluations, so unidimensionality cannot be presumed for a score built from applications either [12].

Nonparametric item response theory, and Mokken scale analysis in particular, was developed for this class of question [13–15]. The monotone homogeneity model requires that the probability of a positive response increases monotonically with the latent trait and that responses are locally independent, and it justifies ordering respondents by their sum score. The more restrictive double monotonicity model additionally requires that the ordering of item difficulties is the same for every respondent, and it is this invariant item ordering that underwrites the interpretation of a scale as a hierarchy in the substantive sense [14, 16, 17]. The two models are not interchangeable, and applied analyses that examine both commonly find the weaker satisfied and the stronger not: all 14 items of a mental well-being scale met the monotone homogeneity model while four violated double monotonicity [18], and for a health-status instrument the item ordering became accurate only after one item was removed [19]. Where the stronger model is not examined, a hierarchical claim rests on evidence that does not carry it.

A second question arises wherever an instrument is administered in more than one language. The Autonomous Province of Bolzano-South Tyrol has a German-speaking majority and an Italian-speaking minority, both served by a single public health system, and survey instruments are administered in parallel language versions. Where translations are literal, measurement equivalence is plausible on linguistic grounds; where translations are culturally adapted, so that the wording of an item differs in more than surface form, equivalence becomes an empirical question. Even for literal translations the assumption is not safe: in a randomised experiment with bilingual respondents, the language of administration alone produced item bias, including in items with minimal wording, and the perceived

quality of the translation did not predict its size [20]. Translation practice in cross-cultural instrument development treats the demonstration of equivalence as evidence about the translation procedure rather than as a formality to be assumed [21, 22]. Differential item functioning (DIF) analysis addresses this question directly: an item functions differentially if respondents at the same level of the latent trait but in different groups have different probabilities of endorsing it [23, 24]. Reliance on significance testing alone is known to flag trivial differences in large samples, which is why effect-size criteria are recommended alongside [25].

The third question is practical. A 16-item block is long for a general population survey and longer still for a survey of health professionals with limited time, and instrument length is a recognised constraint on response quality and participation [26], although a meta-analysis of response rates could not separate the effect of length from that of content and concluded that the choice between instruments should rest on what they measure [27]. Deriving a short form from a longer instrument in a population-representative sample is an established design, and the criteria for such derivation — retained scalability, retained coverage of the item difficulty range, and demonstrated agreement with the full instrument — are well described [28, 29]. The second of these criteria is the one most easily lost: selecting items by their loadings alone favours homogeneous and mutually redundant items and narrows the construct, whereas a selection that protects breadth retains the relations of the instrument to other measures [30].

The present study analysed the internal structure of a 16-item acceptance instrument administered in the February–March 2026 wave of a probabilistic population panel in South Tyrol. The prevalence of acceptance and its sociodemographic correlates are reported separately [31]; the present analysis is concerned with structure and addresses three questions. First, do the 16 applications form an ordered, cumulative scale, and if so, how far does that ordering extend — to groups only, or to individual respondents? Second, do the items function equivalently in the German and the Italian version, given that the Italian version shifts the scope of the application described in three of them and omits a qualifier in two further ones? Third, can the structure be reproduced with a substantially shorter item set suitable for administration in settings where the full instrument is not feasible? A fourth, exploratory question asks whether respondents cluster into distinguishable acceptance types or whether variation along the dimension is purely gradual.

Methods

Study design and sample

Data came from the February–March 2026 wave of "So denkt Südtirol / Così pensa l'Alto Adige", a probabilistic panel maintained by the Provincial Institute of Statistics of the Autonomous Province of Bolzano-South Tyrol (ASTAT). The panel is recruited from the resident adult population with stratification by sex, municipality and age class. The wave was fielded as an online questionnaire; of 1200 panel members invited, 901 responded, a wave participation rate of 75.1%. This figure is not a response rate from the general population, because panel members had already passed an earlier recruitment stage:

the panel recruitment rate was approximately 20%, so that the cumulative response rate is approximately 15% [31]. Poststratification weights were calibrated on sex, age, area of residence and educational level; the weights sum to 408,524 and thus project to the adult resident population rather than being normalised to the sample size. Weighted estimates of acceptance for individual AI applications were previously published in an official ASTAT statistical bulletin, which also included selected multivariable analyses of AI use and general attitudes toward AI in health care [32]. A more comprehensive analysis of these outcomes and their sociodemographic and technology-related correlates is reported separately [31]. Sample characteristics are given in Table 1 and correspond to those reported for the same wave in the prevalence study [31].

Table 1
Characteristics of the sample used in the analyses reported here (N = 901)

Characteristic	Category	Unweighted, n (%)	Weighted, %
Sex	Male	426 (47.3)	50.0
	Female	475 (52.7)	50.0
Age group, years	18–40	287 (31.9)	33.3
	41–60	367 (40.7)	37.6
	61–80	247 (27.4)	29.1
Educational attainment	Lower secondary or vocational	291 (32.3)	33.6
	Upper secondary (Matura)	369 (41.0)	44.6
	University degree or higher	241 (26.7)	21.8
Language group	German	613 (68.0)	68.0
	Italian	213 (23.6)	24.0
	Ladin or other	75 (8.3)	8.0
Technological affinity	Low	31 (3.4)	3.6
	Middle	254 (28.2)	28.4
	High	616 (68.4)	68.1
Frequency of AI use	Never	161 (17.9)	18.5
	Less than once per week	205 (22.8)	23.1
	Approximately once per week	115 (12.8)	12.6
	Several times per week	273 (30.3)	29.7
	Daily	147 (16.3)	16.1

The table is restricted to the variables entering the analyses reported here. Sex, age group and language group are three of the four grouping variables of the differential item functioning analysis; the fourth, area of residence, is dichotomised as the city of Bolzano-Bozen (n = 170) against the remainder of the province (n = 731) and is not tabulated separately. The covariates of the latent class model are the six characteristics listed above together with area of residence. Employment status and the full sociodemographic description of the sample are reported in Table 1 of the prevalence study [31], from which the values above are taken unchanged; the wave, the sample and the poststratification weights are identical.

Weighted percentages are based on poststratification weights reflecting the distribution of sex, age, area of residence and educational attainment in the target population. Technological affinity is the mean of three items on technology use, interest and self-assessed competence, categorised at the cut-points used in the prevalence study; the boundary between the middle and the high category lies at 3.67 inclusive.

Reporting follows the COSMIN reporting guideline for studies on measurement properties, supplemented by STROBE for the descriptive component; the completed checklist, with the section of this manuscript in which each item is addressed, is provided as Additional file 1 [33, 34].

The 16-item acceptance instrument and its two language versions

Respondents were presented with 16 applications of AI in health care and asked, for each, how acceptable or useful they considered the use of AI systems in that area. Response options were "no", "yes", "don't know" and "I am not familiar with this aspect". The applications span administrative functions (appointment scheduling, documentation, electronic health record management, medication reminders, planning and coordination of treatment), monitoring and data analysis (vital-parameter monitoring, analysis of clinical data), diagnostic and therapeutic support (evaluation of diagnostic tests, support in diagnosis, individualised treatment plans, risk assessment and prognosis in chronic disease), direct care and interaction (nursing support, patient information and education through chatbots, robotic surgery), and time-critical decision support (emergencies and triage, urgent decisions). An item-by-item comparison of the two language versions is provided in Additional file 2; the full questionnaire is published as Multimedia Appendix 2 of the prevalence study [31].

The items are not taken from a validated instrument. When the wave was designed, in late 2025, the three Italian-validated instruments for acceptance of AI were either unpublished or available only in preliminary validation, and none existed in a German version suitable for a bilingual general-population survey [3, 31]. The 16 applications were therefore adapted from the corresponding block of a population survey conducted in Germany within the HeReCa panel [6], drafted in German by the Provincial Institute of Statistics together with the Institute of General Practice and Public Health, and translated into Italian by the Provincial Institute of Statistics. Content completeness, face validity and linguistic equivalence were reviewed by one collaborator from each institution; a quantitative content validity assessment and

cognitive pretesting were not carried out [31]. The structural properties of the item set are therefore established here after the fact, which is why they are reported in full rather than summarised.

The questionnaire was administered in parallel German and Italian versions. Three of the 16 items differ from a literal translation in the scope of the application described, and two further items omit a qualifier present in the German version; all five are treated here as culturally adapted. The three shifts in scope are the following: the German item on the evaluation of diagnostic tests, illustrated by x-ray and computed tomography, corresponds to an Italian item restricted to diagnostic images; the German item on patient information and education through chatbots corresponds to an Italian item on automated communication with patients; and the German item on the planning and coordination of treatment pathways between departments and professionals corresponds to an Italian item on coordination between departments or professionals, without the planning component. The two omitted qualifiers are these: the German item on support in drawing up individualised treatment plans corresponds to an Italian item on support in treatment planning, and the German item on nursing support and the performance of nursing measures corresponds to an Italian item on assistance in nursing activities. The four response options are equivalent in both versions. These departures from literal translation motivate the equivalence analysis reported below rather than being incidental to it.

Dichotomisation and treatment of uncertainty responses

Responses were dichotomised into acceptance versus non-acceptance. The main analyses scored both uncertainty categories as non-acceptance, consistent with the index construction used in the prevalence analysis [31]. This choice is substantively defensible, because neither response expresses endorsement, and it preserves the full sample. A sensitivity variant treating both categories as missing was computed; because only 199 respondents gave a substantive answer — "yes" or "no" rather than one of the two uncertainty responses — to every one of the 16 applications, there being no item-level missing data, this variant is reported as a robustness note rather than as an equivalent analysis.

Weighting

Psychometric coefficients were computed unweighted. They are properties of the measurement instrument at the person level rather than population quantities, and the mokken package does not accommodate survey weights. Prevalence statements — item difficulties, class sizes, sum-score distributions — are additionally reported weighted. The separation is empirically inconsequential in these data: the rank correlation between unweighted and weighted item difficulties is exactly 1.000 and the largest difference is 0.6 percentage points, so the ordering that carries the substantive argument does not depend on the weighting.

Mokken scale analysis

Scalability was assessed through the Loevinger coefficients H, H_i and H_{ij} with standard errors. Conventional benchmarks treat H ≥ 0.30 as a weak scale, ≥ 0.40 as moderate and ≥ 0.50 as strong [13, 14]. Monotonicity was checked for each item with rest-score groups of at least 90 respondents, the package default of one tenth of the sample at this sample size, and a minimum reportable violation of 0.03; the crit statistic was evaluated against the conventional warning threshold of 40 and the critical value of 80 [35, 36]. Invariant item ordering was assessed by the manifest invariant item ordering method, with the same rest-score group size and the same minimum reportable violation of 0.03, a nominal alpha of .05 for the significance tests of these violations, through the H^T coefficient and the associated item-elimination procedure; H^T below 0.30 indicates that person ordering by item hierarchy is not supported, values between 0.30 and 0.40 indicate weak, between 0.40 and 0.50 moderate and above 0.50 strong ordering performance. The coefficient is reported throughout for the item set under discussion and not for the reduced set that the elimination procedure leaves behind, since the two differ whenever items are removed — the reported values were obtained with the elimination step disabled, and the elimination itself was run separately to identify which items it removes [17]. Local independence was assessed through the standard conditional-association procedure and, because that procedure is known to be liberal at this sample size, additionally through Yen's Q3 residual correlations, centred on their median, with values above 0.20 treated as indicating dependence [37, 38]. Dimensionality was examined through the automated item selection procedure at lower bounds of 0.30, 0.40 and 0.50.

Two additional checks addressed the stability of the ordering. Over 200 random split halves of the sample, H and the rank correlation of item difficulties between halves were recorded. Across eight subgroups — the two language groups, the two extreme age groups, women and men, and the provincial capital against the remainder of the province — the rank correlation of item difficulties with the overall ordering was computed.

Differential item functioning

For each item, three nested logistic regression models were fitted: the item response on the sum score; on the sum score and the grouping variable; and on the sum score, the grouping variable and their interaction. The increment in McFadden's pseudo-R^2 attributable to the grouping variable was the primary criterion, with 0.035 as the threshold for practical relevance [25]. Likelihood ratio tests for uniform and non-uniform DIF were computed and the resulting p values adjusted by the Benjamini-Hochberg procedure across the 16 items [39]. Language group (German versus Italian) was the primary grouping; area of residence, sex and age (18–40 versus 61–80 years) were secondary. Area of residence was included because language group and urbanity are confounded in South Tyrol (Cramér's V = 0.48), so that an apparent language effect could reflect an urban–rural difference and vice versa. As a robustness check, all analyses were repeated conditioning on the rest score rather than the total score.

A multi-group confirmatory factor analysis of the five attitude items included elsewhere in the questionnaire was considered and rejected, because with five indicators, one of the two factors being measured by two items, a scalar invariance test would have been insufficiently robust [40]; the

differential item functioning analysis of the 16 acceptance items addresses the equivalence question more directly and with greater power.

Derivation and validation of the short form

The search space was exhaustive: all 4368 five-item and all 8008 six-item subsets of the 16 applications were evaluated, so that no candidate was excluded by a prior restriction. The automated selection procedure of Mokken scale analysis is approximate and need not return the optimal partition, and the genetic algorithm proposed to improve it is itself an approximation to a search over all partitions [41]; with 16 items the complete search is feasible directly. The procedure also does not by itself account for sampling fluctuation in the scalability coefficients [42], which is why the selection rule and not merely the selected form was cross-validated. Candidate short forms were assessed on five criteria simultaneously: scalability (H), person ordering (HT), local independence (largest centred Q3), agreement with the 16-item sum score (Spearman rho) and coverage of the difficulty range (the span between the easiest and the most difficult item). Because a short form is intended to describe and compare groups rather than merely to reproduce a total score, a sixth, practical criterion was added: how well the short score separates the acceptance types identified in the latent class analysis, measured as the proportion of correctly ordered pairs drawn from the rejecting and the comprehensively accepting class, and as the rank correlation between short score and class membership.

The decision rule combined these criteria in the following form: among the five-item subsets with a largest centred Q3 of at most 0.20, a rank correlation of at least 0.90 with the 16-item score and a difficulty span of at least 0.50, the form with the highest person ordering was selected. The two constraint values follow the conventional criteria for local dependence and for agreement between a short form and its parent instrument; they were not, however, fixed before the analysis, and the selection is sensitive to the second of them.

The selection procedure, not merely the selected form, was then cross-validated over 200 random split halves. In each repetition the rule was applied afresh in one half and the resulting form evaluated in the other, in both directions, giving 400 selection-evaluation pairs. This yields an estimate outside the sample in which selection took place and quantifies the optimism of the procedure. The distributions of H and H^T are reported as means and 95% ranges. This procedure uses the sample more efficiently than a single split and yields distributions rather than point estimates.

For the selected form, internal consistency was quantified as Cronbach alpha, Molenaar-Sijtsma rho and Guttman lambda-2. Alpha is reported because COSMIN requires a statement on internal consistency [33], and the two additional coefficients because alpha assumes tau-equivalence, which a Mokken scale does not impose; the Molenaar-Sijtsma statistic is the corresponding quantity for nonparametric item response models [43]. The conversion between the short and the full score was estimated by linear regression and, because a regression alone conceals the dispersion behind it, additionally tabulated as the distribution of the 16-item score at each value of the short score.

Latent class analysis

The scale analysis establishes that the applications can be ordered on one dimension; it cannot show whether every respondent evaluates them along that dimension. A latent class analysis was added as an exploratory check for response patterns that a single ordered dimension does not represent. Latent class models with two to six classes were fitted to the dichotomised items on the full sample with 20 random starts per class number. Model selection considered the Bayesian information criterion (BIC), sample-size adjusted BIC, Akaike information criterion, the bootstrapped likelihood ratio test (BLRT) with 500 draws, entropy, the size of the smallest class and substantive interpretability. Stability was assessed through 500 nonparametric bootstrap draws, reported as the distribution of the correlation between bootstrap class profiles and the profiles of the original solution and as the distribution of the smallest class size. A three-class solution was retained as a robustness variant.

Associations with covariates were estimated through a three-step procedure with BCH correction [44] rather than through naive modal assignment. Because the BCH weights derived from the inverse classification matrix are partly negative, standard multinomial routines are not applicable; the weighted multinomial likelihood was maximised directly, following the conception of the corrected third step as a weighted likelihood for clustered data. Since that likelihood is a pseudo-likelihood, standard errors were obtained from a sandwich estimator rather than from the model-based information matrix, which would understate them [45, 46]. The sandwich estimator treats the classification matrix as known and therefore covers the misspecification introduced by the BCH weighting but not the uncertainty carried over from the estimation of the measurement model in step one. That remaining component was quantified separately by a nonparametric bootstrap over all three steps, in which each of 1000 draws re-estimated the four-class model, rebuilt the classification matrix and refitted the covariate model, with classes reordered by mean acceptance in every draw to prevent label switching. A control computation using proportional assignment was run alongside. Covariates were age group, sex, education, language group, area of residence, frequency of AI use and technology affinity. To identify which covariates produce the change in the age coefficient reported below, the same estimator was applied to three nested specifications: age alone, age with AI use and technology affinity, and the full model.

Deriving the classes in one half of the sample and replicating them in the other was considered and rejected: in preparatory analyses the mean profile correlation between halves was 0.824 with a fifth percentile of 0.73, and the smallest class comprised approximately 60 respondents per half, which is insufficient for a four-class solution. Estimation on the full sample with bootstrap assessment of stability was adopted instead.

Software

All analyses were computed in R 4.5.1, using mokken for the scale analysis, poLCA for the latent class models, survey for the design-based estimates and nnet for the multinomial step of the covariate model. The bootstrapped likelihood ratio test is not provided by poLCA and was implemented directly: data sets

were simulated under the model with k − 1 classes, both models were re-estimated on each simulated set with poLCA, and the observed likelihood ratio was referred to the null distribution so obtained. Data preparation, including the recoding of the source variables and the derivation of the technology-affinity categories, was taken unchanged from the analysis code of the prevalence study [31]. Analysis scripts are available as Additional file 3.

Results

Scalability and dimensionality of the 16-item scale

The 16 applications formed a single scale of moderate strength, with a Loevinger coefficient of H = 0.491. No item fell below $H_i = 0.30$, and no item pair showed a negative H_{ij}; the smallest pairwise coefficient was $H_{ij} = 0.248$, between robotic surgery and patient information through chatbots. Item coefficients ranged from 0.393 (patient information through chatbots) to 0.613 (automated medication reminders). Item difficulties and item coefficients are shown in Table 2, and the resulting hierarchy in Fig. 1.

Table 2
The 16 applications: weighted acceptance and scalability coefficients, ordered by acceptance

Item	Application	Weighted acceptance, % (95% CI)	Unweighted, %	H_i (SE)	Short form
E4_SQ014	Automated medication reminders	79.7 (76.9–82.2)	80.0	0.613 (0.036)	●
E4_SQ001	Appointment scheduling and organization	75.0 (72.0–77.7)	75.1	0.481 (0.035)	
E4_SQ015	Management of the electronic health record	67.8 (64.6–70.8)	67.9	0.503 (0.027)	
E4_SQ008	Analysis of clinical data (patterns, anomalies)	66.2 (63.0–69.3)	66.8	0.567 (0.022)	●
E4_SQ016	Planning and coordination of treatment	62.2 (59.0–65.4)	62.4	0.521 (0.022)	●
E4_SQ002	Documentation of findings, visits, or therapy	60.8 (57.5–64.0)	60.8	0.472 (0.023)	
E4_SQ003	Monitoring of vital parameters	54.8 (51.5–58.0)	55.2	0.478 (0.021)	
E4_SQ006	Support in individualized treatment plans	51.1 (47.8–54.4)	51.4	0.505 (0.019)	
E4_SQ005	Support in diagnosis	51.0 (47.7–54.3)	51.3	0.462 (0.020)	
E4_SQ004	Evaluation of diagnostic tests (eg, x-ray, CT)	47.9 (44.6–51.2)	48.2	0.479 (0.019)	
E4_SQ007	Risk assessment and prognosis in chronic disease	45.1 (41.9–48.4)	45.6	0.474 (0.020)	●
E4_SQ013	Surgical procedures using robotic systems	41.0 (37.8–44.3)	41.1	0.422 (0.022)	
E4_SQ012	Nursing support	40.0 (36.8–43.3)	40.1	0.479 (0.020)	
E4_SQ011	Patient information and education (chatbots)	38.3 (35.1–41.6)	38.1	0.393 (0.024)	
E4_SQ009	Support in medical emergencies and triage	33.6 (30.6–36.8)	33.5	0.520 (0.022)	
E4_SQ010	Support in urgent decisions	28.2 (25.3–31.3)	28.2	0.558 (0.025)	●

Note: Overall scale: H = 0.491. Smallest Hij = 0.248 (robotic surgery with patient information). No negative Hij. Item coefficients are unweighted; acceptance percentages are reported both weighted and

unweighted, with a rank correlation of 1.000 between the two and a maximum difference of 0.6 percentage points. ● marks the five items of the selected short form.

The ordering runs from administrative support at the accepted end to time-critical clinical decisions at the least accepted end. Automated medication reminders were accepted by 79.7% (95% CI 76.9–82.2), appointment scheduling by 75.0% (72.0–77.7) and management of the electronic health record by 67.8% (64.6–70.8), whereas support in medical emergencies and triage was accepted by 33.6% (30.6–36.8) and support in urgent decisions by 28.2% (25.3–31.3).

The sum score over the 16 applications had a mean of 8.46 (SD 4.77) and a median of 9, with a weighted mean of 8.43 (SE 0.161). Scalability was stable across split halves. Over 200 random splits, H was 0.490 in one half and 0.491 in the other, with 95% ranges spanning 0.458 to 0.523, and the rank correlation of item difficulties between halves averaged 0.989 (95% range 0.974–0.997).

The automated item selection procedure supported unidimensionality at the conventional criterion. At a lower bound of 0.30 all 16 items were assigned to a single scale. At 0.40 one item remained unassigned, and only at 0.50 did the item set separate into two scales with two unassigned items.

Monotonicity, invariant item ordering and local independence

Monotonicity held throughout. No item showed a significant violation, and the largest crit value was 18, well below the conventional warning threshold of 40.

Invariant item ordering across respondents was not supported. The H^T coefficient was 0.267, below the value at which invariant item ordering is conventionally regarded as supported, and the item-elimination procedure removed four items — analysis of clinical data, robotic surgery, evaluation of diagnostic tests and patient information through chatbots — in order to establish an invariant ordering; for the twelve items that remain the coefficient is 0.303. The value of 0.267 reported here is that of the full 16-item set, which is what the instrument as administered comprises. The consequence for interpretation is specific: the scale supports group comparisons and the use of the sum score as an ordering of respondents on the latent dimension, but the stronger claim that individual respondents can be located by the item hierarchy is not supported for the 16-item instrument. The short form performs substantially better in this respect (see below).

Local independence was assessed by two procedures with divergent results. The conditional-association procedure excluded 10 of the 16 items, a result consistent with the known liberality of that procedure at this sample size and of limited informational value. The Q3 residual correlations, with a median of − 0.074 and a range of − 0.248 to + 0.363, were more informative. Seven item pairs exceeded the criterion of + 0.20 after centering, and one stood clearly apart: support in medical emergencies and triage with support in urgent decisions, at + 0.437. Five of the remaining six involve administrative applications and the sixth links support in diagnosis with individualised treatment plans, so their

dependence reflects shared content rather than item redundancy in the strict sense, and none of them places two items in the selected short form. The pair of support in diagnosis and evaluation of diagnostic tests, anticipated as a candidate on content grounds, was unremarkable at + 0.080 and ranked 27th of the 120 pairs. Only the emergency pair had practical consequences, because both of its items were contained in the seven-item candidate short form. All 120 pairs are listed in Additional file 4.

Measurement equivalence across language groups and further subgroups

The ordering of item difficulties was practically identical in every subgroup examined. Rank correlations with the overall ordering were 0.985 among German-speaking (n = 613) and 0.987 among Italian-speaking respondents (n = 213), 0.965 in the 18–40 age group (n = 287) and 0.962 in the 61–80 age group (n = 247), 0.988 among men (n = 426) and 1.000 among women (n = 475), and 0.984 in the provincial capital (n = 170) and 0.994 in the remainder of the province (n = 731). The smallest of these coefficients was 0.962. Subgroup correlations are tabulated in Additional file 5.

The primary DIF analysis compared the German and the Italian version among the 826 respondents in the two language groups. No item was significant after false discovery rate adjustment, the smallest q value being 0.079, and no item approached the relevance criterion: the largest increment in pseudo-R^2 attributable to language group was 0.0070, against a threshold of 0.035. The three items with the smallest unadjusted p values were risk assessment and prognosis (p = .005; ΔR^2 = 0.007), support in emergencies and triage (p = .015; ΔR^2 = 0.006) and patient information through chatbots (p = .016; ΔR^2 = 0.005). Conditioning on the rest score rather than the total score did not change any of these assessments.

Of the three items whose Italian version shifts the scope of the application described, none reached the relevance criterion: the increment in pseudo-R^2 was 0.0053 for patient information through chatbots, 0.0007 for the evaluation of diagnostic tests and 0.0003 for planning and coordination of treatment. The two items omitting a qualifier were likewise unremarkable (0.0000 and 0.0008). Equivalence therefore holds for precisely those items for which it was least assured on linguistic grounds.

Among the secondary groupings, area of residence was likewise unremarkable, with a largest ΔR^2 of 0.0042 and a smallest q of 0.171. This is relevant for the interpretation of the language result, because language group and urbanity are confounded in the province; the absence of DIF on both variables makes a mutually masking pair of effects implausible. For sex, one item reached q < .05 in the uniform test, management of the electronic health record (q = .003), and a different item in the non-uniform test, analysis of clinical data (q = .008); the largest ΔR^2 was 0.0124.

The age comparison produced the only divergence between the two criteria. Six items reached q < .05 in the uniform test and two in the non-uniform test, with a smallest q of 0.0001. All effect sizes nevertheless remained below the relevance criterion, the largest being 0.0281 against a threshold of 0.035, although this is approximately four times the largest value observed for language group. The

finding is reported here as it stands: the age difference in item functioning is statistically well established in these data and is not dismissed by the adjustment for multiplicity, but its magnitude does not reach the level at which DIF is conventionally treated as consequential for scale use. Summary DIF results for all four groupings appear in Table 3; item-level results are in Additional file 6.

Table 3
Differential item functioning across four groupings

Grouping	n	Largest ΔR^2	Items above 0.035	Items with q < .05, uniform	Items with q < .05, non-uniform	Smallest q
Language group (German/Italian)	826	0.0070	0	0	0	0.079
Area of residence (province/Bolzano)	901	0.0042	0	0	0	0.171
Sex	901	0.0124	0	1	1	0.003
Age (18–40 / 61–80 years)	534	0.0281	0	6	2	0.0001

Note: ΔR^2 is the increment in McFadden's pseudo-R^2 attributable to the grouping variable; the relevance criterion is 0.035. q values are Benjamini-Hochberg adjusted across the 16 items. Language group is the primary analysis; the remaining three are secondary. Item-level results are in Additional file 6.

Derivation and properties of the short form

The seven-item candidate from the preparatory analyses (medication reminders, analysis of clinical data, planning and coordination, individualised treatment plans, risk assessment, emergencies and triage, urgent decisions) achieved H = 0.602 and a rank correlation of 0.932 with the 16-item score, but carried two weaknesses: person ordering was mediocre (H^T = 0.451) and the largest centred Q3 was 0.434, caused by the retained emergency pair.

Exhaustive evaluation of the 28 five- and six-item subsets and the 63 substitution variants identified a five-item form, adopted here as the final short form, that combines high person ordering with local independence. It comprises automated medication reminders, analysis of clinical data, planning and coordination of treatment, risk assessment and prognosis in chronic disease and support in urgent decisions, and thus covers the full difficulty range; relative to the seven-item candidate, individualised treatment plans and emergencies and triage are removed. It achieved H = 0.597, H^T = 0.507, a largest centred Q3 of 0.166 and a rank correlation of 0.903 with the 16-item score. Of the 4368 five-item subsets, 57 satisfied the three constraints of the decision rule, and this form had the highest person ordering among them. Across the whole search space it ranked 103rd of 4368 by person ordering, at the 97.7th percentile, and 99.6th percentile among the 2626 forms with acceptable local independence; no form dominated it on all five criteria simultaneously.

Two features of the search space bear on the interpretation. First, shortening by itself buys little: the median five-item subset reached a person ordering of 0.290 against 0.267 for the full instrument, and 57.2% of all five-item subsets exceeded the full instrument. The value of 0.507 therefore comes from the selection and not from the reduction in length. Second, unconstrained maximisation of person ordering follows redundancy rather than the common dimension: of the 100 subsets with the highest values, 57 retained both emergency items, with a median centred Q3 of 0.381, and the 15 highest retained them without exception. The forms ranked second to fifth by H^T all retained the emergency pair and showed Q3 values between 0.41 and 0.51. Candidate short forms are compared in Table 4, the full search in Additional file 7.

Table 4
Candidate short forms

Form	k	H	H^T	Largest centred Q3	Rho with 16-item score	Difficulty span	Emergency pair retained
Selected five-item form	5	0.597	**0.507**	**0.166**	0.903	0.52	no
Best six-item form	6	0.602	0.503	0.464	0.918	0.52	yes
Seven-item candidate (starting point)	7	0.602	0.451	0.434	0.932	0.52	yes
Without emergencies and triage	6	0.594	0.442	0.199	0.924	0.52	no
Without urgent decisions	6	0.588	0.393	0.207	0.921	0.47	no
Without both emergency items	5	0.580	0.315	0.150	0.894	0.34	no
Without both, robotic surgery added	6	0.542	0.308	0.177	0.919	0.39	no
Without both, nursing support added	6	0.562	0.328	0.170	0.925	0.40	no

Internal consistency of the selected five-item form: Cronbach alpha 0.738, Molenaar-Sijtsma rho 0.763, Guttman lambda-2 0.744; for the 16-item instrument 0.895, 0.907 and 0.897. Cross-validated over 200 split halves, mean (95% range): selected five-item form H 0.597 (0.549–0.644), H^T 0.507 (0.462–0.552); best six-item form H 0.601 (0.560–0.644), H^T 0.503 (0.459–0.548); seven-item candidate H 0.602 (0.567–0.638), H^T 0.451 (0.412–0.489). The rank correlation of item difficulties between halves was 1.000 for all three. Separation of the extreme acceptance types was perfect (1.000) for every form; the

rank correlation with class membership was 0.851, 0.876 and 0.889 respectively. The full search over all five- and six-item subsets and the cross-validation of the selection procedure are in Additional file 7.

The cross-validation of the selection procedure over 400 selection-evaluation pairs gave a mean H^T of 0.507 against 0.453 for the seven-item candidate, with H essentially unchanged. The advantage in person ordering is therefore consistent across both views of the data. Distributions are given in Table 4 and Additional file 7.

The composition of the form is less firmly determined than its properties. Applying the selection rule afresh in each half-sample returned 34 different five-item forms across the 400 selection-evaluation pairs, and the form adopted here was returned once. Two items were selected almost without exception, automated medication reminders in 100% of the half-samples and support in urgent decisions in 99.8%; the three intermediate positions were interchangeable among several near-equivalent candidates, the most frequently selected being appointment scheduling (77.2%), analysis of clinical data (45.2%) and planning and coordination of treatment (40.2%). Appointment scheduling, the third most frequently selected item overall, is not part of the adopted form. The three intermediate items were chosen on content grounds, because analysis of clinical data, planning and coordination of treatment and risk assessment cover the diagnostic and therapeutic domains that the accompanying survey of health professionals addresses, whereas the psychometrically equivalent alternatives built on appointment scheduling and nursing support would leave those domains uncovered. What the data determine are the anchors of the scale; the interior is not determined by them. Selection frequencies for all 16 applications are given in Additional file 7.

The role of the two emergency items admits a clear reading: retaining both is harmful, removing both is worse. Retaining one resolves the dependence, and support in urgent decisions is the one to keep, being the most difficult item in the instrument, carrying the higher item coefficient (H_i = 0.558 versus 0.520) and yielding the better person ordering. Removing both truncates the difficult end of the scale (H^T 0.315, span 0.34), and substituting robotic surgery or nursing support does not restore it. The six variants are compared in Table 4 and in Additional file 7.

Against the practical criterion, every candidate form separated the extreme acceptance types perfectly. The rank correlation between short score and class membership declined from 0.889 for the seven-item candidate to 0.851 for the selected form; this decline is the cost of shortening and was accepted in exchange for the gains in person ordering and local independence.

The choice between this form and the best six-item alternative was made on the local-dependence criterion. This alternative — the seven-item candidate without individualised treatment plans — achieves marginally higher scalability (H = 0.602 against 0.597), a slightly higher correlation with the 16-item score (0.918 against 0.903) and a slightly better rank correlation with class membership (0.876 against 0.851), but it retains both emergency items and therefore leaves the local dependence unresolved (largest centred Q3 0.464 against 0.166), with a lower H^T of 0.503. Since the purpose of the short form is to

order persons in settings where the full instrument cannot be administered, and since a scalability coefficient inflated by redundancy between two near-identical items does not serve that purpose, the five-item form was selected. In the split-half validation the two forms are indistinguishable (H^T 0.507 against 0.503, with almost completely overlapping 95% ranges), so the decision rests on local independence and not on person ordering.

The five-item score has a mean of 2.83 (SD 1.61) and a median of 3 in the unweighted sample and a weighted mean of 2.81 (SE 0.055); it uses the full range, with 12.4% of respondents at 0 and 18.2% at 5. Against the conventional threshold of 15% [47] the score therefore shows no floor effect but a modest ceiling effect, which is what a five-item form covering a wide difficulty range is expected to produce. Internal consistency is adequate for a five-item scale: Cronbach alpha 0.738, Molenaar-Sijtsma rho 0.763 and Guttman lambda-2 0.744, against 0.895, 0.907 and 0.897 for the full instrument. Item-rest correlations range from 0.420 (support in urgent decisions) to 0.593 (analysis of clinical data), and removing any single item lowers alpha, to between 0.656 and 0.721, so every item contributes.

Two conversions between the short and the full score are provided. The regression is score16 = 0.82 + 2.70 × score5 with R^2 = 0.836 and a residual standard error of 1.93 points; the Pearson correlation between the two scores is 0.914 against a Spearman correlation of 0.903. The residual standard error indicates the limit of this conversion: it is adequate for group-level comparison but not for locating individual respondents, since the 16-item scores observed at a given short score have standard deviations between 0.85 and 2.18 points. The distribution of the 16-item score at each value of the short score is given in Additional file 7.

Acceptance typology (exploratory)

Latent class models were fitted for two to six classes. No criterion identified a stopping point. The BIC continued to decline through six classes, placing its optimum at the upper boundary of the range evaluated, and the bootstrapped likelihood ratio test with 500 draws rejected the simpler model at every step, returning p = .002 — the smallest attainable value — for all four adjacent comparisons, with likelihood ratios of 906.4, 271.9, 160.1 and 145.5. The selection of four classes therefore rests on entropy, the size of the smallest class and substantive interpretability rather than on formal model-selection criteria. This is the principal limitation of the typology and is stated as such; the typology is reported as exploratory throughout. Model fit statistics are given in Table 5.

Table 5
Latent class models with two to six classes

Classes	Log-likelihood	Parameters	BIC	aBIC	AIC	Entropy	Smallest class, % (model-based n)
2	−7743	33	15,711	15,606	15,552	0.901	32.7 (295)
3	−7290	50	14,920	14,761	14,680	0.871	18.8 (169)
4	−7154	67	14,764	14,551	14,442	0.859	13.6 (122)
5	−7074	84	14,720	14,453	14,316	0.808	12.6 (114)
6	−7001	101	14,690	14,369	14,204	0.821	11.7 (106)

Note: BIC Bayesian information criterion; aBIC sample-size adjusted BIC; AIC Akaike information criterion. The class size in the last column is model-based, that is the class proportion applied to N = 901; modal assignment yields slightly different counts, which are the ones reported in the text. The bootstrapped likelihood ratio test with 500 draws returned p = .002 for all four adjacent comparisons, the smallest attainable value.

The four-class solution had an entropy of 0.859, a mean maximum posterior probability of 0.919 and diagonal classification probabilities of 0.954, 0.896, 0.919 and 0.937. The classes, ordered by mean acceptance, comprise a consistently rejecting class (13.6% of the model-based distribution; 127 respondents by modal assignment; mean 0.43 of 16 applications accepted), an administratively oriented class (29.7%; 266; 5.62), a broadly accepting class (43.0%; 384; 10.8) and a comprehensively accepting class (13.7%; 124; 15.4). Item-response probabilities by class are shown in Fig. 2 and tabulated in Additional file 8.

The administratively oriented class is the substantively informative one, because it is the only class whose profile departs from a simple shift along the common dimension. Its members endorse appointment scheduling (0.772), medication reminders (0.857) and management of the electronic health record (0.632) at levels approaching those of the broadly accepting class, while endorsing emergencies and triage (0.069) and urgent decisions (0.041) at levels approaching those of the rejecting class. The distinction between administrative and clinically consequential applications, which the item hierarchy expresses as a difference in difficulty, is expressed in this class as a categorical boundary.

Stability under resampling was adequate. Across 500 nonparametric bootstrap draws, all of which converged, the correlation between bootstrap profiles and the original solution averaged 0.930, with a fifth percentile of 0.800 and a minimum of 0.704; the smallest class averaged 13.2% with a fifth percentile of 11.2%. For comparison, the split-half variant examined in the preparatory analyses achieved a profile correlation of only 0.824 (fifth percentile 0.73), which is what motivated estimation on the full sample. In the three-class robustness variant (class sizes 18.8%, 48.9% and 32.3%; entropy 0.871) the two middle classes merge: of the 384 respondents in the broadly accepting class, 220 fall into the

middle and 164 into the upper category of the three-class solution. Bootstrap distributions and the three-class solution are reported in Additional file 9.

Weighting did not shift the typology: weighted class shares were 14.4%, 29.4%, 42.6% and 13.6%, against unweighted shares of 14.1%, 29.5%, 42.6% and 13.8% under modal assignment.

All estimates below use sandwich standard errors; these exceed the model-based ones by 1% to 25%, with a median ratio of 1.11, and change the significance of exactly one of the 39 coefficients — a coefficient not otherwise reported here.

Two covariate findings follow. Language group and area of residence were unassociated with class membership in every class (for the contrast between German and Italian all $p \geq .45$, and $\geq .147$ once the Ladin or other category is included; for area of residence all $p \geq .56$), which extends the null finding reported in the prevalence analysis [31] to the person-centred level.

The second concerns age, where the descriptive and the adjusted picture diverge, though for only one of the three contrasts. Descriptively, respondents aged 61 to 80 years were over-represented at both poles of the typology and under-represented in the administratively oriented middle: they constituted 36.2% of the rejecting class and 32.3% of the comprehensively accepting class, but only 19.9% of the administratively oriented class, against 27.4% in the sample as a whole. In the covariate model with BCH correction, taking the broadly accepting class as reference, respondents aged 61 to 80 had lower odds of membership in the rejecting class rather than the broadly accepting class (OR 0.486; 95% CI 0.264–0.896; $p = .021$) and lower odds of membership in the administratively oriented class (OR 0.375; 0.207–0.678; $p = .001$), with a non-significant tendency towards the comprehensively accepting class (OR 1.709; 0.884–3.306; $p = .111$). Only the first of these reverses the descriptive picture. The second agrees with it, since older respondents were already under-represented in the administratively oriented class, and the third does not change direction.

What produces the reversal was established by estimating three nested models with the same procedure. With age as the only covariate, respondents aged 61 to 80 had higher odds of membership in the rejecting class (OR 1.396; 0.827–2.356; $p = .212$), in line with the descriptive picture. Adding AI use and technology affinity turned the coefficient (OR 0.550; 0.304–0.993; $p = .048$), and the remaining covariates changed it little further (OR 0.486). The corresponding coefficients for the administratively oriented class were 0.563, 0.376 and 0.375 and for the comprehensively accepting class 1.221, 1.693 and 1.710; neither changes sign at any step. The reversal is therefore attributable to digital experience rather than to sociodemographic composition: daily AI use was associated with lower odds of membership in the rejecting class (OR 0.060; 0.018–0.201; $p < .001$) and in the administratively oriented class (OR 0.276; 0.119–0.640; $p = .003$), and high technology affinity operated in the same direction (OR 0.472; 0.290–0.769; $p = .003$ and OR 0.604; 0.373–0.979; $p = .041$).

The bootstrap over all three steps qualifies the age result and is reported for that reason. Of 1000 draws, 991 converged. Its standard errors exceed the sandwich ones by a further 19% at the median and the

model-based ones by 30%, which shows that the uncertainty carried over from the estimation of the measurement model is not negligible. Under the bootstrap percentile intervals neither of the two age coefficients that are significant under the sandwich estimator remains separated from unity: the rejecting class gives 0.279 to 1.399 and the administratively oriented class 0.195 to 1.438. The interval for the comprehensively accepting class, 1.055 to 5.350, does exclude unity, but it is strongly right-skewed and its lower bound lies immediately above it; since the sandwich test of the same coefficient gives p = .111, this exclusion is a property of the percentile method rather than a stronger result, and is not interpreted as one. The direction of the reversal is therefore stable across estimators and across the nested models, but its magnitude is not established. The claim supported by these data is that the descriptive over-representation of older respondents at the rejecting pole does not survive adjustment for digital experience — not that older respondents are, net of that adjustment, less likely to reject. The full covariate model with both sets of standard errors, the nested age models and the bootstrap intervals is provided in Additional file 10.

Sensitivity analysis

Treating both uncertainty categories as missing rather than as non-acceptance left 199 respondents, 22.1% of the sample, with complete data across all 16 items. Scalability in this subsample was substantially higher, H = 0.637 against 0.491, and every item coefficient rose with it, the smallest being 0.518 against 0.393 in the main analysis. Person ordering moved in the opposite direction: H^T fell from 0.267 to 0.240. Neither movement, however, is a property of the scoring rule alone, because the complete-case subsample differs from the full sample in who is analysed as well as in how uncertainty is scored. The two components are therefore separated below.

Estimated item by item, against the respondents who took a position on the item in question, acceptance was higher than in the main analysis for every application, by between 12.9 and 25.2 percentage points with a mean of 20.5, and this increment was only weakly related to item difficulty (rho = − 0.25). The ordering was almost untouched: the rank correlation of item difficulties with the main analysis was 0.985, no item moved more than two places, and the four items at the extremes of the scale held exactly the same ranks. The decision to score uncertainty as non-acceptance thus affects the level of acceptance and not its structure.

Restricting the analysis further to the 199 complete cases adds a second component, smaller but differently directed. Relative to the item-wise estimates these respondents were less accepting of the readily accepted applications, by up to 8.9 percentage points, and more accepting of the least accepted ones, by up to 8.1; the difference was strongly ordered by item difficulty (rho = − 0.89). The result is a compression of the scale: the span between the most and the least accepted application falls from 0.52 in the full sample and 0.48 in the item-wise estimates to 0.32 among the complete cases. That compression accounts for the divergence of the two coefficients, since a narrower difficulty range carries less information for ordering persons, and it means that the combined difference of 4.4 to 25.2 percentage points is not a dichotomisation effect; at the median, 30% of it is attributable to selection.

The subsample is selective. Respondents gave a mean of 4.66 uncertainty responses, and 157 respondents — more than the number of complete cases — declined to take a position on more than half of the applications. The number of substantive answers per item shows where this concentrates and tracks the acceptance hierarchy closely (rho = 0.93): medication reminders and appointment scheduling were answered substantively by 776 and 746 respondents, support in medical emergencies and triage by 551. This variant is accordingly reported as a robustness indication rather than as an alternative estimate. Full results are in Additional file 11.

Discussion

Public acceptance of artificial intelligence in health care is not a set of separate judgements about separate applications but can be characterised as a single ordered dimension. The 16 applications examined here form a moderately strong cumulative scale, and the ordering of that scale is substantive rather than arbitrary: it runs from applications that support administration towards applications that bear on clinical decisions, and within the latter from decisions taken with deliberation towards decisions taken under time pressure. Of the 302 respondents who accepted AI for support in emergencies and triage, 92.1% also accepted it for automated medication reminders, whereas of the 721 who accepted medication reminders only 38.6% accepted it for triage; the corresponding proportions for support in urgent decisions are 95.3% and 33.6%. The practical consequence is that the position of an application in this ordering, rather than its technical sophistication or its degree of automation, predicts how the population will evaluate it.

That the rank ordering of item difficulties is identical in the German- and the Italian-speaking population is the second finding, and it is the one with the widest methodological reach. Rank correlations of item difficulties with the overall ordering did not fall below 0.962 in any subgroup examined, and no item showed differential functioning by language group that approached the conventional relevance criterion. The result is a positive finding rather than the absence of a negative one, and it is strengthened rather than weakened by the translation history of the instrument: three items depart from literal translation in the scope of the application described and two further items omit a qualifier, and equivalence holds for precisely these. Where a translation has been adapted for cultural fit, demonstrated equivalence is evidence about the adequacy of the adaptation procedure and not merely a licence to pool the groups [21, 48]. Because the analysis proceeds item by item, the result also covers the application-level prevalences reported in the prevalence study: where single items rather than a scale score are compared across language groups, item-level differential functioning weighs more heavily, not less [49], and none was found here. Because language group and area of residence are confounded in South Tyrol, and because neither variable produced relevant DIF, the finding is not an artefact of a masked urban–rural contrast. For bilingual health systems and for cross-national research more generally, this supports the administration of the instrument in parallel language versions without stratified scoring [50, 51].

The limits of the hierarchy claim require the same directness as the claim itself. The invariant item ordering that would license the location of individual respondents on the item hierarchy is not supported

for the 16-item instrument (H^T = 0.267), and four items had to be removed before an invariant ordering could be established. The scale is therefore suitable for comparing groups and for scoring respondents on the latent dimension, but it does not support statements about individual respondents derived from the item order. Instruments for which such statements are made reach considerably higher values: a mobility index used in stroke rehabilitation reports 0.87 for the same coefficient and derives item-based start-and-stop rules for individual patients from it [52]. The tutorial and update literature stresses that invariant item ordering has to be investigated separately and cannot be inferred from scalability alone; where that step is omitted, a demonstration of monotone homogeneity can be read as though it established double monotonicity [14, 53]. Reporting it explicitly costs a strong-sounding sentence and preserves the interpretability of the remainder.

The short form addresses this limitation rather than merely shortening the instrument. The five items proposed here achieve markedly better person ordering than the full instrument (H^T = 0.507 versus 0.267) while retaining scalability (H = 0.597) and correlating 0.903 with the 16-item score. That value is not an artefact of selecting and evaluating in the same data: cross-validation of the selection procedure over 400 selection-evaluation pairs gave 0.518 outside the selection sample, an optimism of 0.027. The mechanism is visible in the analysis of the two emergency-related items: their joint retention inflates scalability through redundancy rather than through the common dimension — the Loevinger coefficient responds to properties of the item set that lie outside the requirement it is meant to test [54] — and removing one of them raises person ordering while leaving the difficulty span intact. Removing both, however, truncates the difficult end of the scale and cannot be compensated by substituting other items, which is a reminder that coverage of the difficulty range is a design requirement of a cumulative scale rather than a desirable extra [29].

The ordering itself reproduces a regularity that descriptive surveys have reported across health systems, namely that administrative applications attract more support than applications bearing on clinical decisions [4–10]. The systematic reviews of patient and public attitudes report approval application by application and do not examine the relations among applications [55, 56]. What the present analysis adds is that the regularity is not a set of separate marginal differences but a single ordering that holds within respondents, so that the position of an application in the ordering carries information about how it will be evaluated even where no survey estimate exists for it. The ordering is not the only source of variation in acceptance: a probability-based vignette experiment in Germany found the attributes of the system, above all its reliability and secondarily its transparency, to weigh more heavily than whether the application concerned diagnosis or treatment [57]. The two results are compatible rather than contradictory, because the items analysed here vary the application while leaving the properties of the system unspecified; what the hierarchy orders is the task assigned to the system and not the system itself. Whether the same ordering appears in populations served by health systems organised differently, or in languages more distant from the two examined here, is open: a multinational survey of hospital patients in 43 countries and 26 languages reports levels of trust and concern, but from a nonprobability sample and without an analysis of internal structure [3, 58]; a survey of the Iranian public, conducted in a

language and a health system further removed from those examined here, likewise reports levels of trust and a preference for physician authority rather than any ordering among applications [59].

A concrete illustration of what the ordering adds is provided by proposals to direct AI at the implementation gap in public health, on the argument that the population-level value of the technology lies in automated case finding, outreach and navigation for interventions of established effectiveness rather than in diagnosis [60]. The applications that such a programme combines do not occupy one region of the ordering reported here. Identification of eligible patients from linked records has its closest counterparts in the management of the electronic health record and the analysis of clinical data, in the upper middle of the ordering at 67.8% and 66.2%; the reminders and the scheduling that carry patients through the sequence lie at the accepted end at 79.7% and 75.0%; but the conversational outreach on which such proposals turn has its counterpart in patient information and education through chatbots, at 38.3% the fourteenth of the 16 applications. That the communication of a positive test result is reserved for a clinician in the sequence proposed is consistent with the position of decision-related applications at the least accepted end. What the hierarchy supports is therefore not that such a programme will be accepted or rejected as a whole, but that its components will be evaluated separately and unequally, and that the component carrying most of the burden is not the one the population is most willing to delegate. For the 29.7% of respondents in the administratively oriented class the point is sharper, since for them the boundary between administrative and clinically consequential applications is categorical rather than gradual.

For the instrument itself the practical consequence is a short form that can be carried where 16 items cannot: omnibus surveys with a fixed item budget, panels in which several topics compete for the same questionnaire space, and designs with repeated measurement, where the full block would have to be administered more than once. The relevant comparison in such settings is not between the short form and the full instrument but between the short form and no measurement at all. Where the object of study is the comparison of populations on identical items, the full instrument remains the better choice, and the accompanying survey of health professionals administers all 16 applications for precisely that reason. Because the two anchoring items are empirically determined and the intermediate ones are not, the form can be adapted to a different setting by replacing the intermediate items with applications relevant to that setting, provided the anchors and the coverage of the difficulty range are retained. Among health professionals the determinants of acceptance are described as transparency of the system, training and fit with the clinical workflow [61], so the applications that occupy the intermediate positions are the ones most likely to require substitution.

The exploratory typology adds a qualification to the dimensional picture rather than an alternative to it. Three of the four classes are ordered along the same dimension and differ essentially in level. The administratively oriented class is the exception: its members accept administrative applications at levels close to the broadly accepting class and time-critical applications at levels close to the rejecting class. This is the boundary between administration and clinical consequence expressed categorically rather than as a difference in degree, and it suggests that for approximately three in ten respondents the

distinction is not one of intensity but one of kind. Focus-group work with patients describes the same boundary in words: acceptance is granted to artificial intelligence as a support tool and withheld where it would act as an independent decision-maker [62]. The claim is advanced tentatively, because the class number is not supported by information criteria or by the bootstrapped likelihood ratio test, and the entire typology is reported as exploratory for that reason.

The reversal of the age pattern between the descriptive and the adjusted analysis is of independent interest for survey practice. Older respondents were over-represented at both poles of the typology, a pattern that in a descriptive presentation would support the conclusion that older people are polarised rather than uniformly sceptical. After adjustment for digital experience, education, sex, language and area of residence, the direction of the association reverses: the point estimates place older respondents less often in the rejecting and in the administratively oriented class. The bimodality is thus a compositional effect of the unequal distribution of AI use and technology affinity within the older age group. How far that reversal can be pressed is limited by the estimator: once the uncertainty from the estimation of the measurement model is included through a bootstrap over all three steps, none of the age coefficients is clearly separated from unity. What the data support is the negative statement — the descriptive over-representation at the rejecting pole is not a net age effect — rather than its positive counterpart. Reporting the descriptive pattern without the adjustment would be misleading; suppressing it because the adjusted model points the other way would conceal a genuine feature of the sample. Both are reported here, with the width of the adjusted intervals stated alongside.

Limitations

The 16 applications were adapted from a German population survey rather than drawn from a validated instrument, and no validation study exists for the source items or for the present adaptation. Their content was reviewed for completeness, face validity and linguistic equivalence, but no quantitative content validity assessment and no cognitive pretesting were carried out [31]. Formal content validation would have added an expert panel rating each application for relevance and representativeness, quantified in content validity indices [63], and a review by members of the target population, which is where omissions and unacceptable items are usually identified [64]. What is reported here are therefore the structural properties of an item set whose content coverage has not been formally evaluated: the applications span administrative, monitoring, diagnostic, therapeutic, interactive and time-critical domains, but no claim is made that they exhaust the field, and an application absent from the instrument cannot be placed in the hierarchy.

What is reported here is an evaluation of internal structure and of measurement invariance across the two language versions, not a comprehensive assessment of measurement properties. No comparator instrument was administered in the same wave, so construct validity in the sense of testing hypotheses about expected correlations remains untested, and because the applications were presented once, reliability over repeated measurement and measurement error were not examined. The accompanying survey of health professionals offers the opportunity to close the first of these gaps by carrying an

established instrument alongside. A nine-item German measure of attitudes towards artificial intelligence in work, health care and education, invariant across sex, age, education and employment status, is the least burdensome option [65]; the twenty-item Oxford scale of public attitudes towards artificial intelligence in medicine is longer but medicine-specific, and its own validation reports a correlation of 0.70 with a general artificial intelligence attitude scale against a negligible correlation with vaccine hesitancy, which is the pattern a comparator has to show to be informative [66]. Neither closes the second gap: reliability over repeated measurement and measurement error require a second administration to the same respondents.

The short form was selected and evaluated in the same data, and no procedure applied within a single dataset removes this entirely. Cross-validating the selection procedure bounds the resulting optimism at 0.027 in person ordering, but it does not substitute for external validation. Because the accompanying survey of health professionals administers the full 16-item block, it permits both an external estimate of the properties of the selected form and an independent repetition of the selection itself, so that whether the same five items are returned in a population whose evaluations are formed under different conditions of knowledge and responsibility becomes testable rather than assumed. That population differs from the one studied here in more than its expertise, however, and confirmation in a further general-population sample remains necessary before the composition of the form can be regarded as settled.

The composition of the short form is therefore not uniquely determined, as the selection frequencies reported above show. Only the two anchors of the scale are fixed by the data; the choice among the near-equivalent intermediate items was made on content grounds. A different investigator applying the same rule to the same data could arrive at a different set of three intermediate items with almost identical measurement properties, and any use of the short form should be read with that in mind.

The covariate model rests on a three-step estimator whose uncertainty is only partly captured. The sandwich estimator corrects the pseudo-likelihood misspecification of the third step; the bootstrap over all three steps shows that a further 19% of standard error is attributable to the estimation of the measurement model itself. That the correction introduces a variance component of its own, so that uncorrected standard errors and confidence intervals are optimistic, is an established property of the procedure rather than a feature of these data [67]. The two age coefficients that are significant under the sandwich estimator are no longer separated from unity once that component is included, and the third is separated only by a right-skewed percentile interval whose lower bound lies immediately above unity; the corresponding statements in the Results section are formulated accordingly. The bootstrap draws re-estimated the class model with fewer random starts than the main model for reasons of computing time, which may have inflated the dispersion somewhat; the resulting intervals should therefore be read as an upper bound on the uncertainty rather than as a point estimate of it.

The class number of the latent class analysis is not supported by formal criteria. The BIC declined through six classes without turning, and the bootstrapped likelihood ratio test rejected the simpler model

at every comparison, returning the smallest attainable p value even with 500 draws; neither criterion offers a stopping point within the range evaluated. Four classes were selected on entropy, class size and interpretability, following the stepwise procedure set out for applied users of the method [68]; a systematic review of health applications documents that selection on these grounds is the prevailing practice where the formal criteria do not converge [69]. The typology is reported as exploratory throughout.

Scoring uncertainty responses as non-acceptance is a substantive decision that affects every coefficient reported here. Uncertainty responses were themselves patterned in this survey — they were more frequent among women, among respondents with lower education and among infrequent AI users [31] — so the dichotomisation conflates two response processes that a polytomous model might separate, which is what threshold and multi-process item response models were developed to disentangle [70]. The sensitivity analysis limits the reach of this objection without removing it: the ordering of item difficulties is essentially unchanged under the alternative treatment (rho = 0.985 when acceptance is estimated item by item and 0.968 when the analysis is restricted to complete cases, with no item moving more than two places under either), so the hierarchy does not depend on the decision, but the level of acceptance does, and the alternative subsample of 199 complete cases is too selective to serve as an independent estimate.

Psychometric coefficients are unweighted. The empirical case for this is the rank correlation of 1.000 between weighted and unweighted item difficulties and the maximum difference of 0.6 percentage points, but the coefficients themselves remain sample rather than population quantities.

The DIF analysis by age produced six items significant after adjustment with effect sizes below the relevance criterion. The judgement that these differences are inconsequential rests on a conventional threshold, and a reader who weights statistical significance more heavily would reach a different conclusion. The divergence is not peculiar to these data: in large samples many items can show statistically significant differential functioning while the net effect on decisions taken from the scale remains negligible [71]. The underlying numbers are reported so that this judgement can be reviewed rather than accepted.

The equivalence finding is a demonstration of the absence of relevant differential functioning at the sample sizes available, not a proof that none exists. With 613 German-speaking and 213 Italian-speaking respondents the analysis is adequately powered for the moderate differential functioning that the effect-size criterion is meant to capture, but small effects cannot be excluded [72], the more so because the adjustment for multiplicity that guards against false positives also lowers power against small differences [73]. What the data show is that the largest increment in pseudo-R^2 attributable to language group was a fifth of the relevance criterion and that no item survived adjustment for multiplicity; this makes relevant differential functioning implausible without ruling out differences too small to matter for scale use.

Finally, the study is confined to one region and one wave. The hierarchy reported here is stable across every subgroup examined within that region, but its generalisation to other health systems and other languages is an empirical question that the present data cannot settle.

Conclusions

Public acceptance of artificial intelligence in health care is organised as a cumulative hierarchy ordered by the clinical consequence of the application, and that hierarchy is invariant across language groups, age groups, sexes and areas of residence in a bilingual population. The instrument functions equivalently in a culturally adapted Italian version and in the German original, which supports its use in bilingual health systems without stratified scoring. A five-item short form reproduces the hierarchy with better person ordering than the full instrument and acceptable loss of agreement; it is offered for settings in which a 16-item block cannot be administered, while the full instrument remains preferable wherever populations are to be compared on identical items. Whether the same hierarchy holds among health professionals, whose evaluations are formed under different conditions of knowledge and responsibility, is the question the accompanying survey of the regional health service was designed to answer, and it carries the full item set for that reason.

Abbreviations

aBIC sample-size adjusted Bayesian information criterion · **AI** artificial intelligence · **AIC** Akaike information criterion · **AISP** automated item selection procedure · **ASTAT** Provincial Institute of Statistics of the Autonomous Province of Bolzano-South Tyrol · **BCH** Bolck-Croon-Hagenaars · **BIC** Bayesian information criterion · **BLRT** bootstrapped likelihood ratio test · **CI** confidence interval · **COSMIN** COnsensus-based Standards for the selection of health Measurement INstruments · **CT** computed tomography · **DIF** differential item functioning · **FDR** false discovery rate · **GDPR** General Data Protection Regulation · **H** Loevinger scalability coefficient · **H^T** coefficient of invariant item ordering · **OR** odds ratio · **Q3** Yen's residual correlation statistic · **SABES** Südtiroler Sanitätsbetrieb / Azienda Sanitaria dell'Alto Adige · **SD** standard deviation · **SE** standard error · **STROBE** STrengthening the Reporting of OBservational studies in Epidemiology

Declarations

Clinical trial number: not applicable.

Ethics approval and consent to participate. The study was approved by the Institutional Review Board of Claudiana — College of Health Professions, Bolzano (protocol 04/2026). The survey forms part of the Provincial Statistical Programme 2024–2026, approved by the Provincial Government of South Tyrol (resolution no. 43 of 9 February 2024); ASTAT is responsible for official data collection in the Autonomous Province of Bolzano-Bozen under Regional Law No. 12 of 22 May 1996. Personal data were

processed in accordance with Regulation (EU) 2016/679 and Italian Legislative Decree 196/2003. ASTAT transferred the anonymised dataset in encrypted form to the Institute of General Practice and Public Health for analysis. Participation was voluntary. Informed consent was obtained before the start of the survey by active confirmation via a checkbox in the online questionnaire. Participants received a 20% discount on admission to a regional spa or to one of the South Tyrolean provincial museums as compensation.

Consent for publication. Not applicable.

Availability of data and materials. The data underlying this study were collected by the Provincial Institute of Statistics (ASTAT) within the "So denkt Südtirol" panel and are subject to Regulation (EU) 2016/679 and Italian Legislative Decree 196/2003. Because of these legal restrictions and the terms under which panel participants consented to data use, the individual-level dataset cannot be made publicly available. Anonymised data may be made available by ASTAT and the Institute of General Practice and Public Health upon reasonable request and subject to a data-sharing agreement. The complete analysis code and all result tables are provided as Additional file 3.

Competing interests. The authors declare that they have no competing interests.

Funding. This research received no external funding.

Authors' contributions. Conceptualization: DA, GP, NC, SL, TG, VB, DHvSP, CJW. Methodology: DA, GP, SL, TG, VB, CJW, WW. Software: NC, SL, TG. Validation: SL, DA, WW. Formal analysis: CJW, WW. Investigation: SL. Resources: TG, DHvSP. Data curation: DA. Writing – original draft: CJW, DA. Writing – review and editing: all authors. Supervision: DHvSP, TG. All authors have read and agreed to the submitted version of the manuscript.

Acknowledgements. We thank the participants of the "So denkt Südtirol" panel for taking part in this survey wave.

During the preparation of this manuscript, the authors used a large language model (Claude Opus 5, Anthropic, San Francisco, CA, USA) in August 2026 to draft the manuscript text from the authors' own analysis results, to structure the tables and additional files, and to check internal consistency. All statistical analyses were computed by the authors in R; no analysis was delegated to the model, and the model had no access to the microdata. The authors have reviewed and edited all output, verified all factual statements and citations against primary sources, and take full responsibility for the content of this publication.

References

1. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022;28(1):31–38. doi:10.1038/s41591-021-01614-0

2. Angus DC, Khera R, Lieu TA, Liu VX, Ahmad FS, Anderson B, et al. AI, health, and health care today and tomorrow: the JAMA Summit report on artificial intelligence. JAMA. 2025;334(18):1650–1664. doi:10.1001/jama.2025.18490
3. Wiedermann CJ, Piccoliori G, Hager von Strobele Prainsack D, Ausserhofer D. Conditional acceptance and the optimism–knowledge gap: a scoping review of attitudes and perceptions of artificial intelligence in healthcare in Italy. Med Sci (Basel). 2026;14(2):276. doi:10.3390/medsci14020276
4. Yakar D, Ongena YP, Kwee TC, Haan M. Do people favor artificial intelligence over physicians? A survey among the general population and their view on artificial intelligence in medicine. Value Health. 2022;25(3):374–381. doi:10.1016/j.jval.2021.09.004
5. Fritsch SJ, Blankenheim A, Wahl A, Hetfeld P, Maassen O, Deffge S, et al. Attitudes and perception of artificial intelligence in healthcare: a cross-sectional survey among patients. Digit Health. 2022;8:20552076221116772. doi:10.1177/20552076221116772
6. Negash S, Papon H, Apfelbacher T, Koçman SE, Gundlack J, Manuilova I, et al. Perceptions of the German general population towards implementing artificial intelligence in medical care: a population-based survey. Stud Health Technol Inform. 2025;327:1130–1134. doi:10.3233/SHTI250566
7. Scott IA, Carter SM, Coiera E. Exploring stakeholder attitudes towards AI in clinical practice. BMJ Health Care Inform. 2021;28(1):e100450. doi:10.1136/bmjhci-2021-100450
8. Witkowski K, Dougherty RB, Neely SR. Public perceptions of artificial intelligence in healthcare: ethical concerns and opportunities for patient-centered care. BMC Med Ethics. 2024;25(1):74. doi:10.1186/s12910-024-01066-4
9. Kauttonen J, Rousi R, Alamäki A. Trust and acceptance challenges in the adoption of AI applications in health care: quantitative survey analysis. J Med Internet Res. 2025;27:e65567. doi:10.2196/65567
10. Beets B, Newman TP, Howell EL, Bao L, Yang S. Surveying public perceptions of artificial intelligence in health care in the United States: systematic review. J Med Internet Res. 2023;25:e40337. doi:10.2196/40337
11. Guttman L. A basis for scaling qualitative data. Am Sociol Rev. 1944;9(2):139–150. doi:10.2307/2086306
12. Schepman A, Rodway P. Initial validation of the general attitudes towards artificial intelligence scale. Comput Human Behav Rep. 2020;1:100014. doi:10.1016/j.chbr.2020.100014
13. Mokken RJ. A theory and procedure of scale analysis: with applications in political research. The Hague: Mouton; 1971. doi:10.1515/9783110813203
14. Sijtsma K, van der Ark LA. A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. Br J Math Stat Psychol. 2017;70(1):137–158. doi:10.1111/bmsp.12078
15. van Schuur WH. Mokken scale analysis: between the Guttman scale and parametric item response theory. Polit Anal. 2003;11(2):139–163. doi:10.1093/pan/mpg002

16. Sijtsma K, Meijer RR. A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. Appl Psychol Meas. 1992;16(2):149–157. doi:10.1177/014662169201600204
17. Ligtvoet R, van der Ark LA, te Marvelde JM, Sijtsma K. Investigating an invariant item ordering for polytomously scored items. Educ Psychol Meas. 2010;70(4):578–595. doi:10.1177/0013164409355697
18. Stochl J, Jones PB, Croudace TJ. Mokken scale analysis of mental health and well-being questionnaire item responses: a non-parametric IRT method in empirical research for applied health researchers. BMC Med Res Methodol. 2012;12:74. doi:10.1186/1471-2288-12-74
19. Feng YS, Jiang R, Pickard AS, Kohlmann T. Combining EQ-5D-5L items into a level summary score: demonstrating feasibility using non-parametric item response theory using an international dataset. Qual Life Res. 2022;31(1):11–23. doi:10.1007/s11136-021-02922-1
20. Velez YR, Saavedra Cisneros Á, Gomez J. Language barriers: causal evidence of linguistic item bias in multilingual surveys. Polit Behav. 2024;46(2):1213–1232. doi:10.1007/s11109-023-09869-8
21. Hawkins M, Cheng C, Elsworth GR, Osborne RH. Translation method is validity evidence for construct equivalence: analysis of secondary data routinely collected during translations of the Health Literacy Questionnaire (HLQ). BMC Med Res Methodol. 2020;20:130. doi:10.1186/s12874-020-00962-8
22. Gjersing L, Caplehorn JR, Clausen T. Cross-cultural adaptation of research instruments: language, setting, time and statistical considerations. BMC Med Res Methodol. 2010;10:13. doi:10.1186/1471-2288-10-13
23. Swaminathan H, Rogers HJ. Detecting differential item functioning using logistic regression procedures. J Educ Meas. 1990;27(4):361–370. doi:10.1111/j.1745-3984.1990.tb00754.x
24. Teresi JA, Fleishman JA. Differential item functioning and health assessment. Qual Life Res. 2007;16 Suppl 1:33–42. doi:10.1007/s11136-007-9184-6
25. Jodoin MG, Gierl MJ. Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. Appl Meas Educ. 2001;14(4):329–349. doi:10.1207/S15324818AME1404_2
26. Galesic M, Bosnjak M. Effects of questionnaire length on participation and indicators of response quality in a web survey. Public Opin Q. 2009;73(2):349–360. doi:10.1093/poq/nfp031
27. Rolstad S, Adler J, Rydén A. Response burden and questionnaire length: is shorter better? A review and meta-analysis. Value Health. 2011;14(8):1101–1108. doi:10.1016/j.jval.2011.06.003
28. Petrowski K, Paul S, Zenger M, Brähler E. An ultra-short screening version of the Recalled Parental Rearing Behavior questionnaire (FEE-US) and its factor structure in a representative German sample. BMC Med Res Methodol. 2012;12:169. doi:10.1186/1471-2288-12-169
29. Smith GT, McCarthy DM, Anderson KG. On the sins of short-form development. Psychol Assess. 2000;12(1):102–111. doi:10.1037/1040-3590.12.1.102

30. Passarelli M, Casetta L, Rizzi L, Chiorri C, Cassina F, Voi S, et al. Short and sweet: comparing strategies for the reduction of questionnaires on self-criticism and social safeness while preserving construct validity. Int J Psychol. 2024;59(6):1234–1244. doi:10.1002/ijop.13249
31. Ausserhofer D, Colletti N, Lombardo S, Gärtner T, Rina P, Barbieri V, et al. Public attitudes toward and acceptance of artificial intelligence in health care in a multilingual alpine region: population-based cross-sectional survey. JMIR Preprints. Preprint posted online August 11, 2026. doi:10.2196/preprints.109238
32. Statistisches Landesinstitut der Autonomen Provinz Bozen-Südtirol (ASTAT). Nutzung von KI und Meinungen zur KI in der Gesundheitsversorgung – Februar 2026 / Uso dell'IA e opinioni sull'IA nell'assistenza sanitaria – febbraio 2026. Astat Info Nr. 21. Bozen/Bolzano: ASTAT; 2026.
33. Gagnier JJ, Lai J, Mokkink LB, Terwee CB. COSMIN reporting guideline for studies on measurement properties of patient-reported outcome measures. Qual Life Res. 2021;30(8):2197–2218. doi:10.1007/s11136-021-02822-4
34. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet. 2007;370(9596):1453–1457. doi:10.1016/S0140-6736(07)61602-X
35. van der Ark LA. Mokken scale analysis in R. J Stat Softw. 2007;20(11):1–19. doi:10.18637/jss.v020.i11
36. van der Ark LA. New developments in Mokken scale analysis in R. J Stat Softw. 2012;48(5):1–27. doi:10.18637/jss.v048.i05
37. Yen WM. Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Appl Psychol Meas. 1984;8(2):125–145. doi:10.1177/014662168400800201
38. Christensen KB, Makransky G, Horton M. Critical values for Yen's Q3: identification of local dependence in the Rasch model using residual correlations. Appl Psychol Meas. 2017;41(3):178–194. doi:10.1177/0146621616677520
39. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol. 1995;57(1):289–300. doi:10.1111/j.2517-6161.1995.tb02031.x
40. Meade AW, Bauer DJ. Power and precision in confirmatory factor analytic tests of measurement invariance. Struct Equ Modeling. 2007;14(4):611–635. doi:10.1080/10705510701575461
41. Straat JH, van der Ark LA, Sijtsma K. Comparing optimization algorithms for item selection in Mokken scale analysis. J Classif. 2013;30(1):75–99. doi:10.1007/s00357-013-9122-y
42. Koopman L, Zijlstra BJH, van der Ark LA. A two-step, test-guided Mokken scale analysis, for nonclustered and clustered data. Qual Life Res. 2022;31(1):25–36. doi:10.1007/s11136-021-02840-2
43. Sijtsma K, Molenaar IW. Reliability of test scores in nonparametric item response theory. Psychometrika. 1987;52(1):79–97. doi:10.1007/BF02293957

44. Vermunt JK. Latent class modeling with covariates: two improved three-step approaches. Polit Anal. 2010;18(4):450–469. doi:10.1093/pan/mpq025
45. Bolck A, Croon M, Hagenaars J. Estimating latent structure models with categorical variables: one-step versus three-step estimators. Polit Anal. 2004;12(1):3–27. doi:10.1093/pan/mph001
46. Bakk Z, Tekle FB, Vermunt JK. Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. Sociol Methodol. 2013;43(1):272–311. doi:10.1177/0081175012470644
47. Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60(1):34–42. doi:10.1016/j.jclinepi.2006.03.012
48. Petersen MA, Groenvold M, Bjorner JB, Aaronson N, Conroy T, Cull A, et al. Use of differential item functioning analysis to assess the equivalence of translations of a questionnaire. Qual Life Res. 2003;12(4):373–385. doi:10.1023/A:1023488915557
49. Bjorner JB, Kreiner S, Ware JE, Damsgaard MT, Bech P. Differential item functioning in the Danish translation of the SF-36. J Clin Epidemiol. 1998;51(11):1189–1202. doi:10.1016/S0895-4356(98)00111-5
50. Heinz A, Sischka PE, Catunda C, Cosma A, García-Moya I, Lyyra N, et al. Item response theory and differential test functioning analysis of the HBSC-Symptom-Checklist across 46 countries. BMC Med Res Methodol. 2022;22:253. doi:10.1186/s12874-022-01698-3
51. Arthurs E, Steele RJ, Hudson M, Baron M, Thombs BD. Are scores on English and French versions of the PHQ-9 comparable? An assessment of differential item functioning. PLoS One. 2012;7(12):e52028. doi:10.1371/journal.pone.0052028
52. Roorda LD, Green JR, Houwink A, Bagley PJ, Smith J, Molenaar IW, et al. Item hierarchy-based analysis of the Rivermead Mobility Index resulted in improved interpretation and enabled faster scoring in patients undergoing rehabilitation after stroke. Arch Phys Med Rehabil. 2012;93(6):1091–1096. doi:10.1016/j.apmr.2011.12.021
53. Sijtsma K, Meijer RR, van der Ark LA. Mokken scale analysis as time goes by: an update for scaling practitioners. Pers Individ Dif. 2011;50(1):31–37. doi:10.1016/j.paid.2010.08.016
54. Roskam EE, van den Wollenberg AL, Jansen PGW. The Mokken scale: a critical discussion. Appl Psychol Meas. 1986;10(3):265–277. doi:10.1177/014662168601000305
55. Young AT, Amara D, Bhattacharya A, Wei ML. Patient and general public attitudes towards clinical artificial intelligence: a mixed methods systematic review. Lancet Digit Health. 2021;3(9):e599–e611. doi:10.1016/S2589-7500(21)00132-1
56. Vo V, Chen G, Aquino YSJ, Carter SM, Do QN, Woode ME. Multi-stakeholder preferences for the use of artificial intelligence in healthcare: a systematic review and thematic analysis. Soc Sci Med. 2023;338:116357. doi:10.1016/j.socscimed.2023.116357
57. Kühne S, Jacobsen J, Legewie N, Dollmann J. Attitudes toward AI usage in patient health care: evidence from a population survey vignette experiment. J Med Internet Res. 2025;27:e70179.

doi:10.2196/70179

58. Busch F, Hoffmann L, Xu L, Zhang LJ, Hu B, García-Juárez I, et al. Multinational attitudes toward AI in health care and diagnostics among hospital patients. JAMA Netw Open. 2025;8(6):e2514452. doi:10.1001/jamanetworkopen.2025.14452
59. Maleki MS, Sahebi L, Shahvari Z, Samadi S. Attitudes of the Iranian public toward the clinical use of artificial intelligence in medicine: a cross-sectional survey. Patient Prefer Adherence. 2026;20:563494. doi:10.2147/PPA.S563494
60. Beckman AL, Chokshi DA. From breakthrough to follow-through—a public health agenda for AI. JAMA. Published online August 13, 2026. doi:10.1001/jama.2026.16748
61. Tun HM, Rahman HA, Naing L, Malik OA. Trust in artificial intelligence-based clinical decision support systems among health care workers: systematic review. J Med Internet Res. 2025;27:e69678. doi:10.2196/69678
62. Gundlack J, Thiel C, Negash S, Buch C, Apfelbacher T, Denny K, et al. Patients' perceptions of artificial intelligence acceptance, challenges, and use in medical care: qualitative study. J Med Internet Res. 2025;27:e70487. doi:10.2196/70487
63. Almanasreh E, Moles R, Chen TF. Evaluation of methods used for estimating content validity. Res Social Adm Pharm. 2019;15(2):214–221. doi:10.1016/j.sapharm.2018.03.066
64. Connell J, Carlton J, Grundy A, Taylor Buck E, Keetharuth AD, Ricketts T, et al. The importance of content and face validity in instrument development: lessons learnt from service users when developing the Recovering Quality of Life measure (ReQoL). Qual Life Res. 2018;27(7):1893–1902. doi:10.1007/s11136-018-1847-y
65. Gnambs T, Stein JP, Appel M, Griese F, Zinn S. An economical measure of attitudes towards artificial intelligence in work, healthcare, and education (ATTARI-WHE). Comput Human Behav Artif Hum. 2024;3:100106. doi:10.1016/j.chbah.2024.100106
66. Kantor J, Morrison M, Ko J, Semenov YR, Fortgang RG. Development and validation of the Oxford Attitudes to Artificial Intelligence in Medicine Scale. JAAD Int. 2026;26:113–121. doi:10.1016/j.jdin.2026.04.005
67. Bakk Z, Oberski DL, Vermunt JK. Relating latent class assignments to external variables: standard errors for correct inference. Polit Anal. 2014;22(4):520–540. doi:10.1093/pan/mpu003
68. Sinha P, Calfee CS, Delucchi KL. Practitioner's guide to latent class analysis: methodological considerations and common pitfalls. Crit Care Med. 2021;49(1):e63–e79. doi:10.1097/CCM.0000000000004710
69. Zhou M, Thayer WM, Bridges JFP. Using latent class analysis to model preference heterogeneity in health: a systematic review. Pharmacoeconomics. 2018;36(2):175–187. doi:10.1007/s40273-017-0575-4
70. Böckenholt U, Meiser T. Response style analysis with threshold and multi-process IRT models: a review and tutorial. Br J Math Stat Psychol. 2017;70(1):159–181. doi:10.1111/bmsp.12086

71. Stark S, Chernyshenko OS, Drasgow F. Examining the effects of differential item (functioning and differential) test functioning on selection decisions: when are statistically significant effects practically important? J Appl Psychol. 2004;89(3):497–508. doi:10.1037/0021-9010.89.3.497
72. Scott NW, Fayers PM, Aaronson NK, Bottomley A, de Graeff A, Groenvold M, et al. A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. J Clin Epidemiol. 2009;62(3):288–295. doi:10.1016/j.jclinepi.2008.06.003
73. Kim J, Oshima TC. Effect of multiple testing adjustment in differential item functioning detection. Educ Psychol Meas. 2013;73(3):458–470. doi:10.1177/0013164412467033

Figures

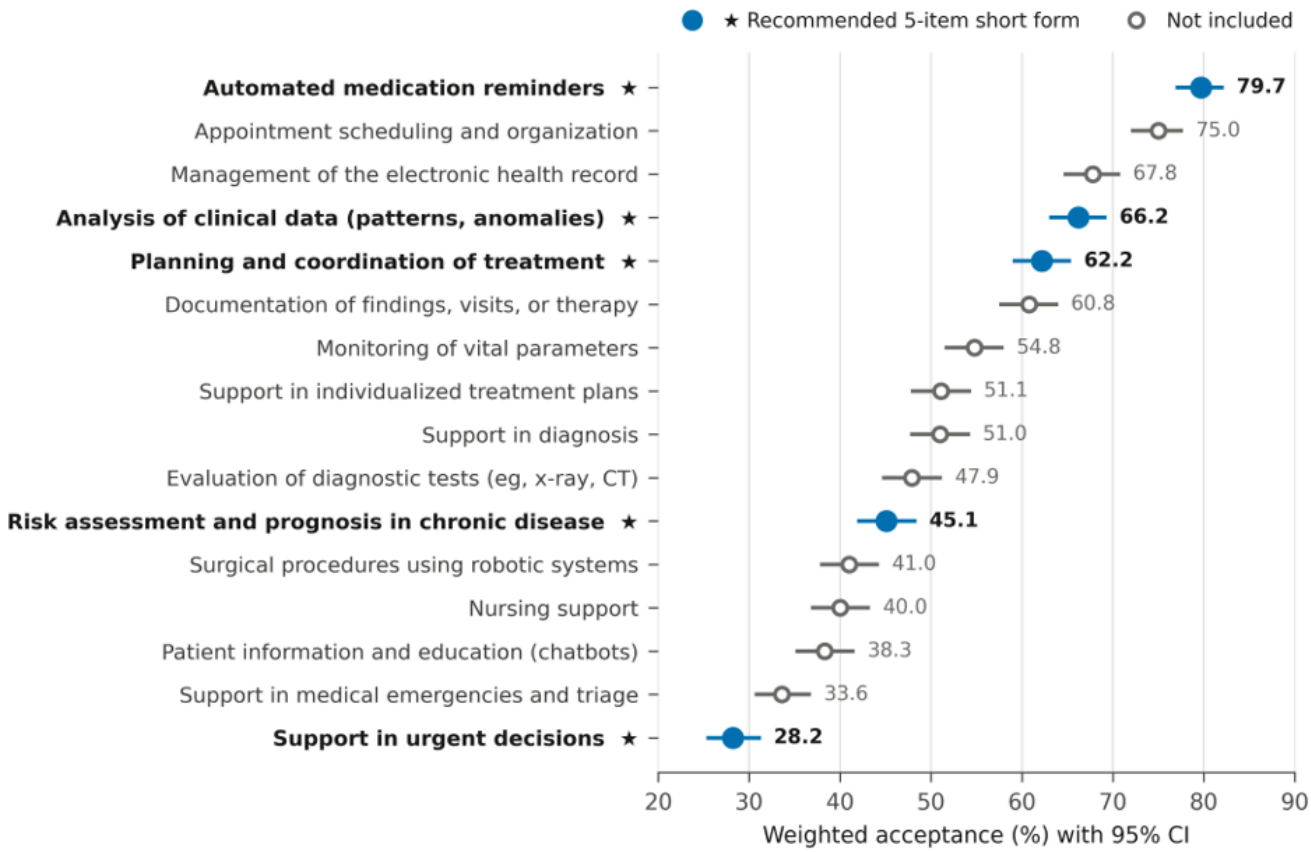


Figure 1

The acceptance hierarchy of 16 artificial intelligence applications in health care. Weighted proportions of respondents rating the application as acceptable, with 95% confidence intervals, ordered from most to least accepted (N = 901). Filled markers and starred labels identify the five items of the selected short form. The ordering runs from administrative support at the accepted end to time-critical clinical decision

support at the least accepted end and was practically identical in all subgroups examined (Spearman rho ≥ 0.962).

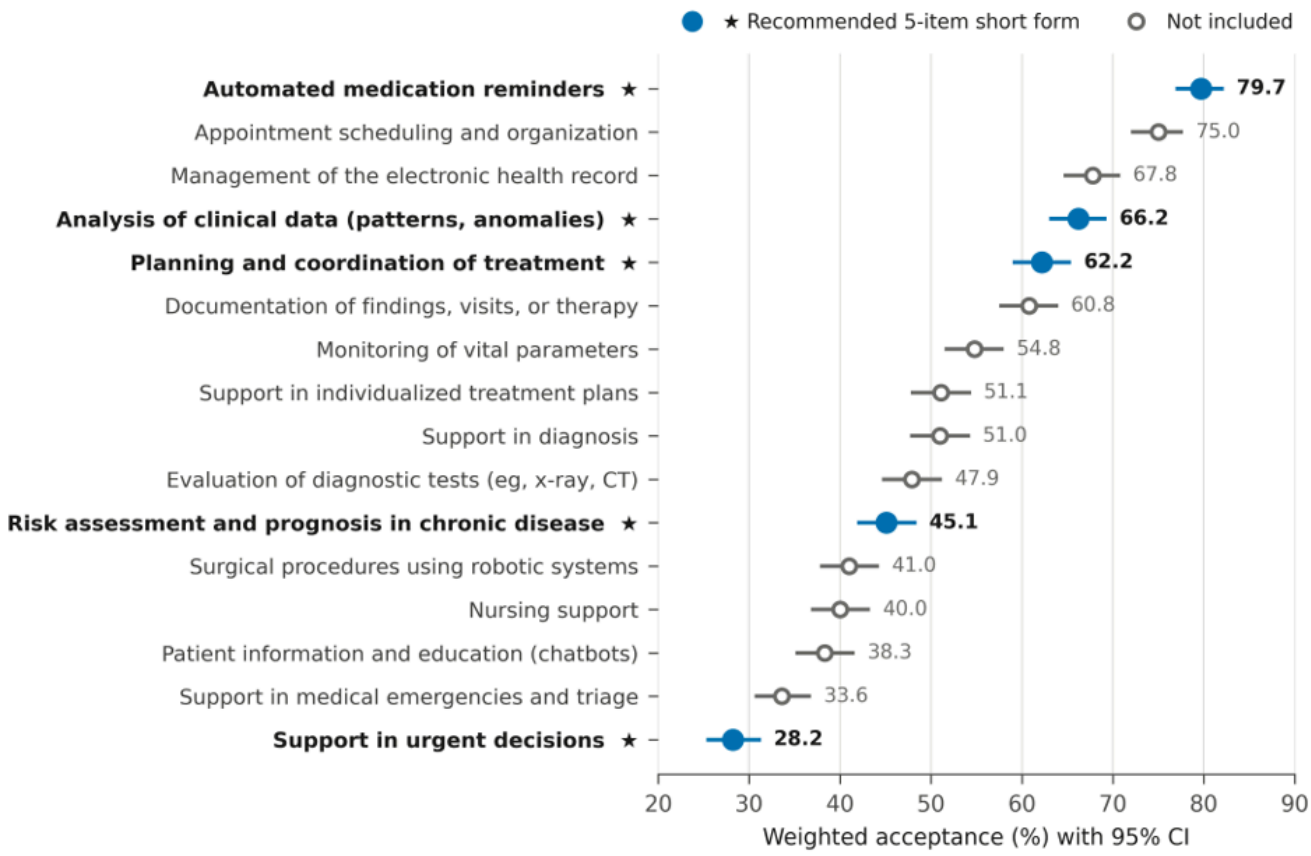


Figure 1

The acceptance hierarchy of 16 artificial intelligence applications in health care. Weighted proportions of respondents rating the application as acceptable, with 95% confidence intervals, ordered from most to least accepted (N = 901). Filled markers and starred labels identify the five items of the selected short form. The ordering runs from administrative support at the accepted end to time-critical clinical decision support at the least accepted end and was practically identical in all subgroups examined (Spearman rho ≥ 0.962).

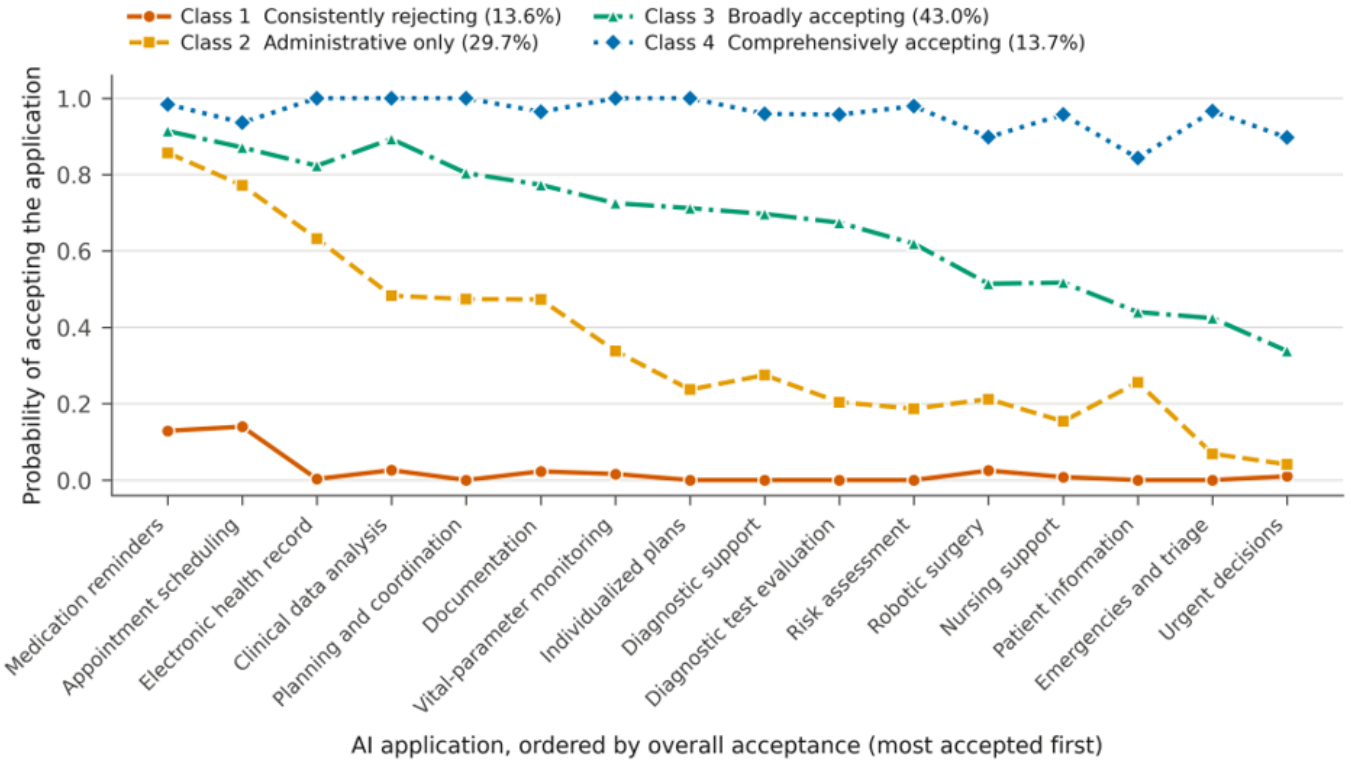


Figure 2

Item-response probabilities by latent class. Probability of accepting each application within each of the four latent classes, with applications ordered by overall acceptance (most accepted first). Class shares are the model-based proportions. Three classes differ essentially in level along the common dimension; the administratively oriented class departs from it, endorsing administrative applications at levels close to the broadly accepting class and time-critical applications at levels close to the rejecting class. The typology is exploratory: the class number is supported neither by information criteria nor by the bootstrapped likelihood ratio test.

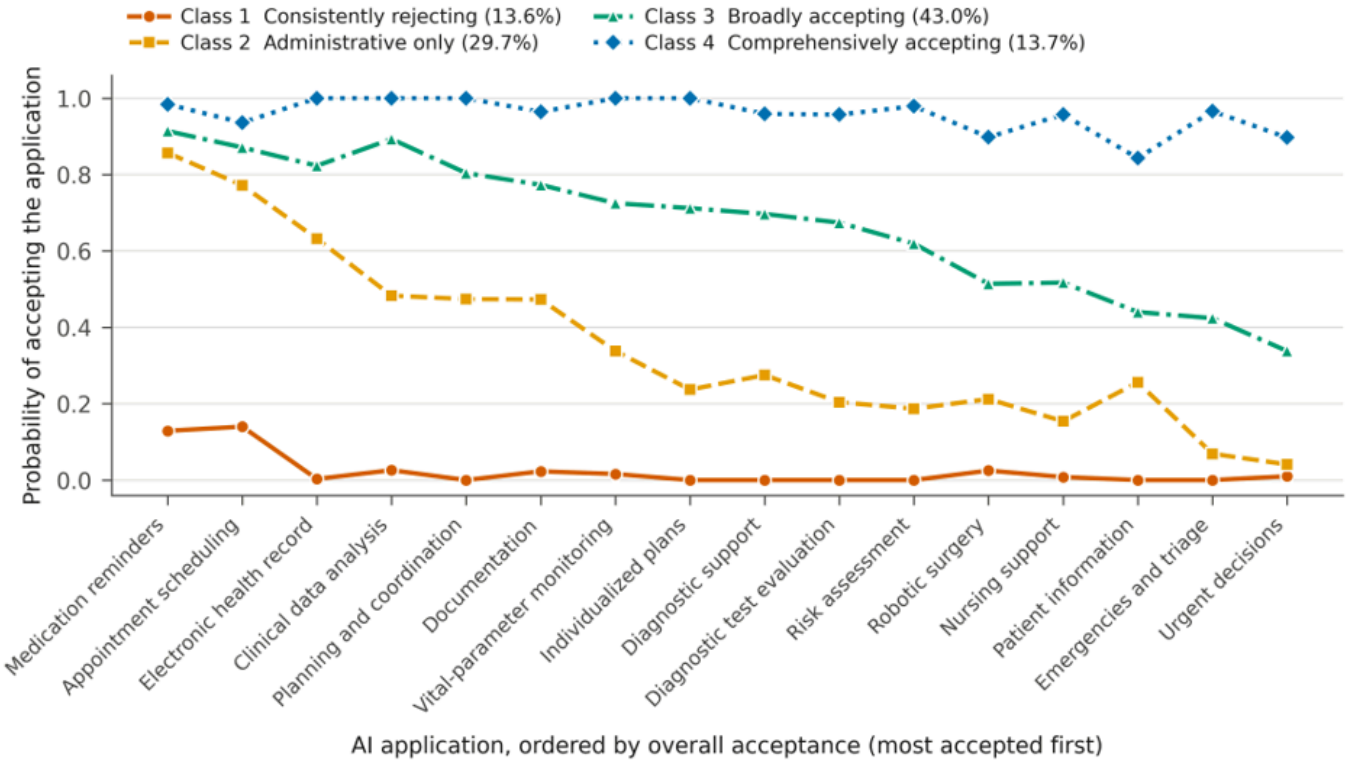


Figure 2

Item-response probabilities by latent class. Probability of accepting each application within each of the four latent classes, with applications ordered by overall acceptance (most accepted first). Class shares are the model-based proportions. Three classes differ essentially in level along the common dimension; the administratively oriented class departs from it, endorsing administrative applications at levels close to the broadly accepting class and time-critical applications at levels close to the rejecting class. The typology is exploratory: the class number is supported neither by information criteria nor by the bootstrapped likelihood ratio test.

Supplementary Files

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

- AdditionalFile11260809.docx
- AdditionalFile11260809.docx
- AdditionalFile01260809.docx
- AdditionalFile01260809.docx
- AdditionalFile02260809.docx

- AdditionalFile02260809.docx
- AdditionalFile08260809.docx
- AdditionalFile08260809.docx
- AdditionalFile09260809.docx
- AdditionalFile09260809.docx
- AdditionalFile03260809.docx
- AdditionalFile03260809.docx
- AdditionalFile04260809.docx
- AdditionalFile04260809.docx
- AdditionalFile05260809.docx
- AdditionalFile05260809.docx
- AdditionalFile06260809.docx
- AdditionalFile06260809.docx
- AdditionalFile07260809.docx
- AdditionalFile07260809.docx
- AdditionalFile03ScriptsandResults.zip
- AdditionalFile03ScriptsandResults.zip
- AdditionalFile10260809.docx
- AdditionalFile10260809.docx
- Additionalfiles.docx
- Additionalfiles.docx