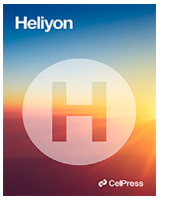
Heliyon 9 (2023) e13074

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

APPRAISE-RS: Automated, updated, participatory, and personalized treatment recommender systems based on GRADE methodology

Beatriz López [a,*], Oscar Raya [a], Evgenia Baykova [b], Marc Saez [c,d], David Rigau [e], Ruth Cunill [f], Sacramento Mayoral [b], Carme Carrion [g], Domènec Serrano [b], Xavier Castells [h]

[a] Control Engineering and Intelligent Systems (eXiT), University of Girona, Spain
[b] Institute of Health Care (ICS-IAS), Girona, Spain
[c] Research Group on Statistics, Econometrics and Health, University of Girona, Spain
[d] CIBER of Epidemiology and Public Health (CIBERESP), Madrid, Spain
[e] Cochrane Iberoamerica, Barcelona, Spain
[f] Sant Joan de Deu-Numancia Health Park, Barcelona, Spain
[g] Health Lab Research Group, Universitat Oberta de Catalunya, Spain
[h] TransLab Research Group, Dept. of Medical Sciences, University of Girona, Spain

ARTICLE INFO

Keywords:
Treatment recommender systems
Evidence-based medicine
Meta-analysis
Attention deficit hyperactivity disorder

ABSTRACT

Purpose: Clinical practice guidelines (CPGs) have become fundamental tools for evidence-based medicine (EBM). However, CPG suffer from several limitations, including obsolescence, lack of applicability to many patients, and limited patient participation. This paper presents APPRAISE-RS, which is a methodology that we developed to overcome these limitations by automating, extending, and iterating the methodology that is most commonly used for building CPGs: the GRADE methodology.

Method: APPRAISE-RS relies on updated information from clinical studies and adapts and automates the GRADE methodology to generate treatment recommendations. APPRAISE-RS provides personalized recommendations because they are based on the patient's individual characteristics. Moreover, both patients and clinicians express their personal preferences for treatment outcomes which are considered when making the recommendation (participatory). Rule-based system approaches are used to manage heuristic knowledge.

Results: APPRAISE-RS has been implemented for attention deficit hyperactivity disorder (ADHD) and tested experimentally on 28 simulated patients. The resulting recommender system (APPRAISE-RS/TDApp) shows a higher degree of treatment personalization and patient participation than CPGs, while recommending the most frequent interventions in the largest body of evidence in the literature (EBM). Moreover, a comparison of the results with four blinded psychiatrist prescriptions supports the validation of the proposal.

* Corresponding author.
E-mail addresses: beatriz.lopez@udg.edu (B. López), oscar.raya@udg.edu (O. Raya), evgenia.baykova@ias.cat (E. Baykova), marc.saez@udg.edu (M. Saez), DRigau@santpau.cat (D. Rigau), rcunill@pssjd.org (R. Cunill), sacramento.mayoral@ias.cat (S. Mayoral), mcarrionr@uoc.edu (C. Carrion), domenec.serrano@ias.cat (D. Serrano), xavier.castells@udg.edu (X. Castells).

https://doi.org/10.1016/j.heliyon.2023.e13074
Received 29 November 2022; Received in revised form 4 January 2023; Accepted 16 January 2023

Available online 24 January 2023

Conclusions: APPRAISE-RS is a valid methodology to build recommender systems that manage updated, personalized and participatory recommendations, which, in the case of ADHD includes at least one intervention that is identical or very similar to other drugs prescribed by psychiatrists.

1. Introduction

As medical knowledge has continued to grow, clinical practice guidelines (CPGs) have become fundamental tools to bring evidence-based medicine to practice [1]. The implementation of a CPG is complex and requires compilations of recommendations derived from a comprehensive, systematic review of published work on the particular health condition that the CPG is designed for. One of the keys to the success of CPGs is the rigorous methodology that is used to build them, the most suitable of which is that of the Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group [2].

Artificial intelligence (AI), in general, and machine learning methods, in particular, have recently been used to support CPG development and maintenance. For example, [3] use a causal model to selected terms and guide the finding of patterns related to a given disease from publications in PubMed by using a semantic-distance metric. In [4], a rule-based system is proposed to formalize the knowledge required to identify the quality of evidence, which depends on the amount and independence of the systematic reviews and the experts' opinions.

Most of the uses of AI for CPGs focus on the CPGs' obsolescence [5]. Meanwhile, there are other important limitations that compromise their validity, such as the lack of personalization (i.e., the existence of recommendations addressed to subgroups of patients [2], particularly to those with mild or moderate symptom severity [6] or with comorbidities [7]), and participation (the patient participation in CPG development [8]). These limitations may explain why non-adherence to the CPGs' recommendations can be as high as 65% [9,10]). This leads to high rates of patients who fail to receive the most suitable care according to the evidence available [11].

AI-based clinical decision support systems can provide solutions for treatment personalization [12] or information to patients about their health condition [13], which is an essential step in patient participation. Nevertheless, no previous work has addressed the weakness of CPGs from a holistic perspective or generated automated, updated, personalized, and participatory treatment recommendations using a well-established methodology.

This work presents a methodology, which is called APPRAISE-RS (Automated ParticiPatoRy And personalized trEatment Recommender System), to develop recommender systems that provide automated, updated, participatory, and personalized treatments. APPRAISE-RS uses rule-based systems to recommend therapies based on the GRADE heuristic [2](automation). Once a patient's demographic and clinical data is known, APPRAISE-RS gathers information from the medical literature (updated knowledge) based on the patient's characteristics (personalization), and on the patient's and clinician's preferences from a list of treatment outcomes (participatory medicine). Evidence from all possible treatments for a given patient is summarized for the clinicians and the recommendations are ranked in order of their suitability. Thus, at the point of care, updated, personalized, and participatory therapy recommendations ranked according to their suitability are provided automatically. Like CPG, APPRAISE-RS recommendations should help and not replace clinical judgement.

In this study, the APPRAISE-RS methodology has been implemented in the field of attention deficit hyperactivity disorder (ADHD) to experimentally test its feasibility. ADHD is a significant health problem and, although more than one medication has proven to be efficacious, recommendations in existing CPGs are often conflicting.

The rest of this paper is structured as follows. In Section 2, we review the related work. In Section 3, we describe the methodology that we developed. In Section 4, we present the implementation of the proposed method in ADHD. In Section 5, we experimentally demonstrate the validity of this approach for 28 simulated patients. Finally, we present the discussion and conclusion in Sections 6 and 7.

2. Related work

The context of this research is analyzed in terms of automated, updated, personalized, and participatory approaches.

First, several systems automate the implementation of CPGs recommendations to help clinicians in treatment decision-making [14]. For example, [15] is a system that supports general practitioners to provide the best dyslipidemia treatment according to the European Guidelines for CVD prevention. Similarly, [12] presents an implementation of the hyperlipidemia treatment guidelines ATP III (Adult Treatment Panel III). Nevertheless, the validity of such systems depends on guideline updates because the medical evidence is continuously changing. Therefore, these systems quickly become obsolete. Conversely, the APPRAISE-RS approach tackles obsolescence by using the latest medical evidence available in the literature.

Second, the guideline updating process is mainly addressed by automatically updating new medical evidence from publications. For example, [3] uses a knowledge model to retrieve publications related to a given disease from PubMed. Meanwhile, a rule-based system to automatically identify medical evidence in published studies is proposed in [4,16]. The APPRAISE-RS method that is presented in this paper assumes that recovery of the relevant publications is performed manually in a regular basis from a system alert mechanism and that a structured database exists that contains all of the published studies required to generate recommendations. Though the use of a manual approach could pose some concerns regarding applicability concerns, it provides robustness to the

recommendations formulated and we expect to overcome this limitation in a near future with the great developments of deep learning (DL) and other works as [4,16].

Third, regarding treatment personalization, clinical decision support systems for guiding treatment decisions use to focus on copying with multimorbidity. Interestingly, [17] present a clinical decision system based on a fuzzy multicriteria decision model to recommend multiple therapies to patients with type 2 diabetes. [18] combine different CPGs by using a semantic integration framework based on transaction logic and temporal constraints. [19] also analyses the interactions of CPGs to handle interactions with drugs to treat diabetes by using a conceptual model of CPGs with reasoning capabilities regarding different CPGs use cases.

In some approaches, treatment recommendations are based on historical data from previously treated patients (e.g., [20], [21] or [22]). For example, [22], used patient information to weigh up the different therapies that are available, according to the outcome to previous, similar patients. [23] dealt with patients with diabetes but focused on complex health conditions and comorbidities to predict the optimal combination of medicines to minimize the likelihood of early re-admission stemming from exacerbations. The authors proposed using random forest and Bayesian network approaches to determine the optimal medication. [24] used machine learning methods, particularly sequence learning, to predict the next medication (i.e., a medication recommended after an initial medication). In contrast, our approach focuses on a single intervention and dose or a combination of interventions at a given time point; however, a sequence of multiple interventions could be of future interest.

Other types of recommender systems use continuous data gathered from the patient over a period of time to make treatment prescription easier (e.g., [25]). Similarly, dose recommenders, such as the Pepper system [26], support the routine treatment of chronic patients. In general, dose recommender systems differ from our approach because APPRAISE-RS focuses on the choice of intervention and dose, rather than the dose alone. Moreover, dose recommenders are designed for continuous use during treatment, while APPRAISE-RS focuses on treatment selection.

Finally, participatory medicine approaches usually focus on providing patients with information on their health issue [27], including advanced information and communication technology (ICT) tools as games [13], in order to encourage collaboration among patients health professionals to partner in determining the treatment. Involving the patient in the decision-making process, patient compliance with the selected intervention is likely to be high, and improve treatment adherence. Informed patients can express preferences on their treatment that can be taken into account by clinical decision support tools when providing a treatment recommendation. For example, [28] incorporated preferences by using a first-order approach to mitigate the decision-making process in patients with multiple comorbidities that causes treatment interactions. [29] proposed gathering patient preferences in two dimensions: qualitative (preferences on every day life habits, as exercise or food intake) and quantitative (preferences regarding treatment types, as amount of doses per day, pills or injectable, etc.). A multi-agent mechanism was used to deal with preferences, and the final treatment was obtained by the use of multi-criteria decision making. These previous works on preferences are complex to implement, and the elicitation of preferences may require the operation of a knowledge engineering. In the case of APPRAISE-RS, patient preferences are used to generate medical evidence that will be used to make treatment recommendations for the patient; no other personnel that the clinician and the patient is required. Moreover, preferences in APPRAISE-RS can be expressed by both, clinicians and patients.

3. Methodology

APPRAISE-RS is a recommender system methodology that automates, adapts, extends, and iterates the GRADE methodology [2] to formulate automated, up-to-date, participatory, and personalized treatment recommendations from an updated database of clinical studies. Fig. 1 shows the various steps of the APPRAISE-RS methodology.

The first step is to gather a patient's basal demographic and clinical data, and also a predefined list of potential therapeutic preferences (i.e., desired improvements stemming from the intervention, such as symptom severity and undesired adverse effects to avoid) with a rating from 1 (not at all important) to 9 (very important). Both patient and clinician preferences are gathered concurrently.

In the second step, the patient's basal data, together with a pre-existing clinical studies database, is used to build a dataset of relevant clinical studies related to their characteristics (filtering). Their critical preferences (those over a given score) are combined in a pool of preferences.

The third step is to generate scientific evidence and to ascertain its quality for each intervention and preference pooled. This is done by using the relevant studies database and applying meta-analysis techniques to combine the results of the studies on each intervention and preference. Therefore, a pooled result for each intervention and preference is generated, which is followed by a quality assessment (rated from high to low) of the evidence generated.

The fourth step assesses the benefit-risk relationship of each intervention, as well as its quality. This involves two parallel stages. On the one hand, a benefit-risk analysis is performed for each intervention by weighting the results for each preference, yielding one of five possible judgments: favorable to the intervention, probably favorable to the intervention, neither favorable nor unfavorable to the intervention, probably favorable to the placebo, and favorable to the placebo. On the other hand, a quality analysis of this judgment is performed, from high to low.

Finally, a clinical recommendation is generated from the benefit-risk relationship of each intervention and its quality. This can be strongly in favor, weakly in favor, weakly against, or strongly against its use. Thus, a list of treatment recommendations is obtained that fits the particular personalized data of the patient, and which has taken into account both the patient's and the clinician's preferences (participatory).

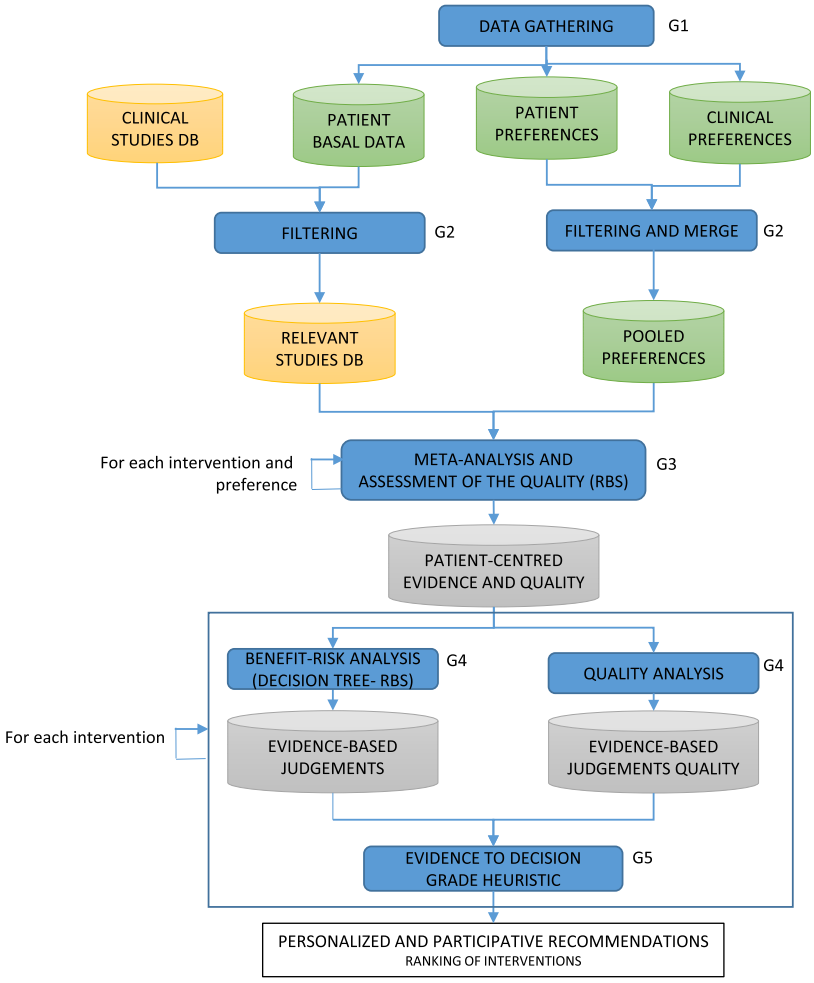


Fig. 1. APPRAISE-RS methodology overview. RBS: rule-based system. G: steps of the GRADE methodology that are used and iterated.

In the point-of-care scenario, the methodology will require from a first visiting appointment in which the patient receives information about their clinical condition, in order to be informed. Next, in a second visit, the patient and clinician express their preferences and obtain a treatment recommendation sorted according to their relevance. The patient and the clinician will select together the final intervention from the recommended ones.

The remainder of this section details the methodological steps, beginning with a substantive explanation of the GRADE method. Note that the GRADE methodology is repeated several times to achieve the final recommendations: first for each intervention and preference (to generate scientific evidences by means of meta-analysis, steps G2 and G3 in Fig. 1); and next for each intervention (to formulate treatment recommendations, steps G4 and G5 in Fig. 1).

3.1. Background to the GRADE methodology

The GRADE methodology requires a panel of experts, who derive clinical recommendations from scientific evidence using the following steps: 1) formulate a clear clinical question with PICO format (i.e., Patient, Intervention, Comparison, Outcome); 2) define the relevance of the outcomes; 3) carry out a systematic review with meta-analysis that answers the PICO question; 4) determine the certainty of the effects as a function of risk bias, inconsistency, whether the evidence is indirect, imprecision of the results, and publication bias; 5) weigh up the results on efficacy outcomes against safety; 6) calculate the economic cost if it is considered to be an important factor in decision-making; 7) evaluate the impact on inequity; 8) determine the acceptability of the intervention; and 9) analyze the feasibility of implementing the intervention [2]. Applying these steps will enable the generation of treatment recommendations, whose strength may be “weak” or “strong”.

APPRAISE-RS adapts and automatizes steps 1-5 of the GRADE method because the remaining steps refer to public health policy issues. First, the patient’s characteristics, and the patient’s and clinician’s preferences are used to formulate the clinical question (i.e., the data gathering step in Fig. 1, labeled with G1). Second, studies of interest are selected (i.e., the filtering step in Fig. 1 labeled with G2). Third, for each studied intervention, preference, or outcome of interest, the effect size is calculated using meta-analysis

Table 1
Structure of the APPRAISE-RS studies database.

Block	Kind of data
Study design ($d_1, ..., d_D$)	Demographic data Comorbidities Counter-indications Drug anamnesis Intervention anamnesis
Sample size ($NPlacebo, NIntervention$)	Number of patients randomized placebo Number of patients randomized intervention
Intervention (I)	Pharmacological treatment
Study results ($o_1, ..., o_R$)	Efficacy results Safety results Acceptability outcome
Risk of bias ($r_1, ..., r_R$)	Risk of bias of efficacy results Risk of bias of safety results Risk of bias of acceptability outcome

techniques and its quality is determined (G3 label in the Figure). Fourth, the assessment of the benefit-risk relationship and its quality are computed (G4). Finally, the treatment recommendation is formulated (G5).

It is worthy to observe that APPRAISE-RS deals with an acceptability preference, but its meaning differs from step 8 of the GRADE approach. APPRAISE-RS uses a quantitative definition of acceptability: an intervention is acceptable if the % of dropouts is lower than in the control group, which translates into a relative risk (RR) < 1,00. In contrast, GRADE's step 8 uses a qualitative definition (the distribution of the benefits, harms and costs). In the step 8 of the GRADE methodology there is not a cut-off (RR < 1.00) that differentiates between acceptable and unacceptable interventions. Furthermore, as GRADE's step 8 relates to public health policy issues, the intervention cost is considered when appraising its acceptability for the health system.

3.2. Clinical studies dataset

The clinical studies dataset contains updated information on publications from trusted sources. It compiles the results of randomized clinical trials (RCTs) to investigate the efficacy and safety of pharmacological interventions for patients with particular diseases. Table 1 shows the design and structure of the database. The data are organized in five blocks: study design, sample size, intervention, study results, and risk of bias. The study design information includes the patient data used as inclusion or exclusion criteria, which is denoted as d_i. Sample size includes the number of patients included in the RCT and distinguishes the number of patients that received a placebo from those who were treated with the intervention. Intervention refers to the pharmacological treatment given to the subjects, which is denoted as I. The study results (o_j) involve efficacy outcomes (i.e., improvement of the clinical manifestations and the repercussions of the disease), safety (i.e., negative unwanted consequences resulting from the administration of a particular treatment), and acceptability of the interventions (i.e., all cause- treatment discontinuations) for the specific study group. A total of R study results is considered. For example, in the case of ADHD, the efficacy outcomes are symptom severity, clinical global impression, prevention of drug use disorder, drug consumption, accidents, academic performance, and quality of life. The safety outcomes are insomnia, drowsiness, dizziness, dry mouth, tics, seizures, vomiting, syncope, and discontinuation due to adverse events. Finally, the risk of bias block includes information about the bias arising from the randomization process, allocation concealment, blinding of patients and investigators, data incompleteness, study reporting, and other sources of deviations [30]. There is one risk of bias per outcome. Therefore, a study of the database is formally noted as a tuple $\langle d_1, ..., d_D$, *NPlacebo, NIntervention,* $I, o_1, ..., o_R, r_1, ..., r_R\rangle$.

The function $kind\_of(o_i) \in$ *{efficacy, safety, acceptability}* is defined to distinguish between the different kind of outcomes. Furthermore, study results whose origins are of a continuous nature (i.e. symptom severity improvement defined in a numerical, continuous scale) are distinguished from those that are binary in nature (i.e. the number of patients experiencing insomnia). The study results of a continuous nature are registered in the database with standardized mean differences (SMDs)[1]; outcomes of binary origin are counters. All of the results that are stored in the dataset are numeric; however, the source of information differs. To distinguish between them, the function $origin(o_i) \in \{continuous, binary\}$ is defined.

Entries in the clinical studies' database relate to the intervention vs. placebo comparison for each study. Therefore, two separate database entries are provided for studies with multiple comparisons (e.g., a three-group study comparing two different pharmacological interventions against one placebo group): one per pharmacological intervention.

The results generated by the methodology depend on the quality of the dataset. Some machine learning tools, such as text mining, can generate a dataset from articles in the literature [32]. However, despite considerable advances in DL for natural language

[1] The computation of the SMD from the data usually provided in the literature can be carried out by following [31].

processing, there are barriers to implementing these approaches in real practice [32]. Therefore, an accurate revision by experts in the disease field is mandatory. This is the reason why APPRAISE-RS requires such a carefully curated dataset as a pre-requisite.

3.3. Data gathering

This step is related to the formalization of the query topic following the GRADE methodology. The data gathered includes the patient's basal information (i.e., demographic and clinical characteristics of the patient) $b_1, ..., b_D$, as well as the patient's and clinician's preferences, $p^p_1, ..., p^p_R$ and $p^c_1, ..., p^c_R$ correspondingly. Regarding basal information, every variable b_i matches the corresponding study design variable d_i in the clinical studies dataset (see Table 1). The preferences are related to the outcomes o_j that the users wish to achieve or avoid. Patients and clinicians must express their preferences by rating the relevance of each corresponding p^j_i variable in accordance with their values and experience. Both patients and clinicians rate each of the available outcomes using a Likert scale [33] that ranges from 1 (not at all important) to 9 (very important). Acceptability preference is always included as a safeguard by default because it balances efficacy and safety [34].

3.4. Study and preference filtering

Given the patient's basal data ($b_1, ..., b_D$), and the patient's ($p^p_1, ..., p^p_R$) and clinician's ($p^c_1, ..., p^c_R$) preferences, the filtering step involves selecting the relevant studies from the clinical studies dataset. First, the patient's demographic and clinical characteristics $b_1, ..., b_D$ are used as a filter to identify the RCTs in the database in which the patient might have participated. This can be carried out because the patient's characteristics are cross-linked with the inclusion/exclusion criteria of the RCTs stored in the clinical studies database (d_i). As a result, a set of relevant studies $RS = \{S_1, ...S_k\}$ is selected for every $d_j \in S_i$, $d_j = b_j$. This first step makes APPRAISE-RS a personalized recommendation methodology because retrieved studies are related to the patient's particularities. For example, in the case of ADHD, if the age of the patient is 7 years old (basal data), then up to 134 studies could be selected from the 264 entries available in the database. A joint retrieval of values 7 for age and man for genre would yield 132 studies.

APPRAISE-RS also analyzes the preferences that the patient and/or clinician consider to be relevant. All of the preferences that are rated ≥ 7 are deemed "critical" and are retained for further consideration. The filtered preferences from the patient and the clinician are merged, resulting in a pool of preferences $p_1, ..., p_Q$, with $Q \leq R$, where $p_j \geq 7$ for all p_j. If both the patient and the clinician select the same preference, then the maximum score is kept. This second filtering step makes APPRAISE-RS recommendations participatory because the user's preferences are taken into account.

3.5. Meta-analysis and quality assessment

In this step, APPRAISE-RS generates patient-centered evidence by means of meta-analysis and assessment of its quality. Meta-analysis combines the results of different RCTs addressing the same question, in our case, the same preference and intervention. For example, and according to Table 2, there are 15 studies (publications) investigating the effect of the pharmacological intervention "Atomoxetine High" on the preference "Somnolence". Each study can show a different result or outcome (e.g. number of patients experiencing "Somnolence"). Meta-analysis is a statistical method that allows for calculating the averaged effect with its corresponding 95% confidence interval, and an estimate of the statistical heterogeneity, which measures the between-study inconsistency in the results. The quality of the meta-analysis results expresses the confidence we have on the averaged effect calculated and it is measured with a 4-point scale ranging from "very low" quality (1 point) to "high" quality (4 points).

All the steps required for performing the meta-analysis are provided in Algorithm 1. First, given the retrieved studies $RS = \{S_1, ...S_k\}$, each of which related to a given intervention $I_1, ..I_k$ (eventually with $I_i = I_j$), a set of candidate interventions $C = \bigcup_{\forall i \in RS}\{I_i\}$ is obtained (step 2 of Algorithm 1).

Next, for every intervention $I_j \in C$ and preference p_i in the pooled preferences $p_1, ..., p_Q$, a subset of studies $RS^{I_j, p_i} \subset RS$ is obtained, such that for every study $S_k \in RS^{I_j, p_i}$ the intervention is the same (I_j) and has an outcome $o_k = p_i$ registered (step 3 to 11 of Algorithm 1). For example, Table 2 shows 51 different studies when the preference "symptom severity" is entered. For the preference "avoid somnolence", up to 85 studies emerge from 105 possibilities. A meta-analysis is performed for each RS^{I_j, p_i} (step 12 of Algorithm 1). Meta-analysis is a method that combines the results from different studies (i.e., RCT) on a similar topic (intervention and preference). This leads to a single estimated outcome and a measure of the statistical heterogeneity, which provides information on the statistical consistency of the pooled result or outcome [35,36].

For preferences related to binary-origin outcomes, APPRAISE-RS first calculates the RR using the number of patients randomized to each intervention, and the number of patients experiencing the event (i.e. somnolence) in each intervention (NPlacebo, Nintervention). Furthermore, the studies contain information on the SMDs of preferences related to continuous-origin outcomes. APPRAISE-RS combines the RRs or SMDs of studies investigating the same intervention and outcome/preference using a random effects model [37] by means of the inverse variance method [38], which results in a pooled effect size (e) and confidence interval (CI). Statistical heterogeneity is determined by calculating the I^2 index [37]. The results of a meta-analysis ma^i_j regarding an intervention I_i and a preference p_j is then a vector of three components $ma^i_j = [e^i_j, CI^i_j, I2^i_j]$.

APPRAISE-RS meta-analysis results are complemented by an assessment of their quality; that is, the extent to which one can be confident that the estimated effects are correct. Quality assessment is obtained by using a rule-based system (RBS) (step 13 and 14 of Algorithm 1). All of the qualities start with a quality score of 4, which is the highest. Next, several conditions are analyzed based on the GRADE method, as follows: risk of bias, heterogeneity, precision of effect estimates, directness of evidence, and publication

Table 2
Example of retrieved studies for ADHD according to two basal data (7 year-old, boy), and two preferences (symptom severity, somnolence). First column: Doses are linked to medicines with an underscore symbol. "+" means combination of treatments. Second column: total retrieved studies. Third and fourth columns: studies retrieved that consider the individual preferences.

Intervention	Retrieved	Symptom severity	Somnolence
Atomoxetine_High	15	15	15
Methylphenidate_High	20	8	16
Atomoxetine_Low	1	1	1
Modafinil_Low	2	1	2
Methylphenidate_Low	15	2	9
Bupropion_High	2	0	2
Dexmethylphenidate_High	4	0	4
Dexmethylphenidate_Low	1	0	1
Lisdexamfetamine_High	4	2	3
Lisdexamfetamine_Low	4	1	2
Modafinil_High	4	4	4
Guanfacine_Low	4	3	4
Guanfacine_High	8	8	8
Mixed amfetamine salts_Low	8	1	4
Mixed amfetamine salts_High	2	0	1
Dexmethylphenidate_Low	1	0	1
Clonidine_High	1	1	1
Clonidine_Low	3	1	3
Methylphenidate_High + Clonidine_Low	1	0	1
Pindolol	1	0	0
Selegiline	1	0	0
Viloxazine	3	3	3
Total	**105**	**51**	**85**

Algorithm 1 Meta-analysis and assessment of its quality.

Require: $RS = \{S_i, ...S_k\}$, $pooled\,Preferences = p_1, ...p_Q$
Ensure: $patientEvidence\&Quality =$
$\langle (I_1, \{(p_1, ma_1^1, q_1^1), ..., (p_Q, ma_Q^1, q_Q^1)\}),$
$...,$
$(I_\Theta \{(p_1, ma_1^\Theta, q_1^\Theta), ..., (p_Q, ma_Q^\Theta, q_Q^\Theta)\})\rangle$

$patientEvidence\&Quality \leftarrow null$
$C \leftarrow \bigcup\{I_i\}$, where $I_i \in S_i$ ▷ Observe $|\,C\,| = \Theta$
for $I \in C$ **do**
  $interventionEvidence\&Quality \leftarrow null$
  for $p \in pooled\,Preferences$ **do**
    $RS^{I,p} \leftarrow \{S \mid S \in RS, I \in S, p \in S\}$
    if $p \sim$ dichotomous outcomes **then**
      for $S \in RS^{I,p}$ **do**
        computeRR(p,S)
      end for
    end if
    $ma \leftarrow metaAnalysis(I, p, RS^{I,p})$
    $q \leftarrow qualityAssessmentRBS(I, p, ma, RS^{I,p})$
    $append((p, ma, q), interventionEvidence\&Quality)$
  end for
  $append((I, interventionEvidence\&Quality),$
    $patientEvidence\&Quality)$
end for

bias [30]. The risk of bias analysis is performed by limiting the meta-analysis to those studies whose risk of bias is deemed to be low, comparing the pooled effect with that of the primary analysis. Heterogeneity is assessed using the I^2 index obtained from the meta-analysis [37]. Precision is determined by the extent to which 95% CI (from the meta-analysis) of the calculated effect overlaps with the null effect (0.00 and 1.00 for SMD and RR, respectively). Directness is not applicable in APPRAISE-RS because all of the

Assessment of heterogeneity
IF the I^2 index ranges from 30 to 49%
THEN subtract 1 point from the quality score.
IF the I^2 index $\geq$ 50%
THEN subtract 2 points from the quality score

Fig. 2. Example of the conditions checked for the meta-analysis quality assessment.

Table 3
Quality ratings from score values.

Score	Rating
4	High
3	Moderate
2	Low
1	Very low

Algorithm 2 Benefit-risk judgement and quality analysis.

Require: $patientEvidence\&Quality = \langle (I_1, \{(p_1, ma_1^1, q_1^1), ..., (p_Q, ma_Q^1, q_Q^1)\}), ..., (I_\Theta, \{(p_1, ma_1^\Theta, \rho_1^\Theta), ..., (p_Q, ma_Q^\Theta, \rho_Q^\Theta)\})\rangle$
Ensure: $judgement\&Quality = \langle (I_1, j_1, q_1), ...(I_\Theta, j_\Theta, q_\Theta)\rangle$
1: $judgement\&Quality \leftarrow null$
2: **for** $I_i \in patientEvidence\&Quality$ **do**
3: $j_i \leftarrow RBS(\{(p_1, ma_1^i), ..., (p_Q, ma_Q^i)\})$ ▷ Remind: $ma_j^i = [e_j^i, CI_j^i, I2_j^i]$
4: $\rho_i \leftarrow min(q_1^i, ..., q_S^i)$
5: $append((I_i, j_i, q_i), judgement\&Quality)$
6: **end for**

studies are retrieved using an exact match from the patient's data. For example, if the patient is an adolescent, then no studies on adults are considered; therefore, the meta-analysis results apply to the patients of interest in a straightforward way. The risk of publication bias is evaluated by comparing the effect of the smallest vs the largest studies. The Egger test is performed when there are at least 10 studies [37].

Fig. 2 shows some examples of the conditions tested (see https://caleta.udg.edu/git/eXiT_Research_Group/APPRAISE-RS-RBS for full details) using APPRAISE-RS. If a condition holds, then the corresponding points are subtracted from the score. The minimum score is 1. Finally, the scores are transformed to the quality ratings, according to the mapping in Table 3. The final assessment may result in high-, moderate-, low-, or very low-quality ratings.

For each intervention I_i, the meta-analysis and quality assessment steps provide a set of evidence and qualities for each pooled preference $(I_i, \{(p_1, ma_1^i, q_1^i), ..., (p_Q, ma_Q^i, q_Q^i)\})$ (step 15 of Algorithm 1). Because the evidence (and its quality) is grounded in the filtered studies and selected preferences, it is patient-centered and personalized.

3.6. Benefit-risk judgement and quality analysis

This step involves summarizing the information on all preferences for a single intervention $(I_i, \{(p_1, ma_1^i, q_1^i), ..., (p_Q, ma_Q^i, q_Q^i)\})$ into a benefit-risk judgement (j_i) and then assessing its quality (ρ_i), (I_i, j_i, ρ_i), as shown in Algorithm 2.

The judgment involves weighting the results on the efficacy outcomes against safety outcomes at an intervention-level (step 2 of Algorithm 2). To perform this judgement, APPRAISE-RS uses a second RBS that was developed from a decision tree provided by EBM experts (step 3 of Algorithm 2). Fig. 3 shows three examples of rules derived from the decision tree (see https://caleta.udg.edu/git/eXiT_Research_Group/APPRAISE-RS-RBS for the full set of rules). As proposed by the GRADE heuristic, the benefit-risk judgment j_i of an intervention I_i can be: "favorable to the intervention", "probably favorable to the intervention", "neither favorable nor unfavorable", "probably favorable to the placebo", or "favorable to the placebo." Moreover, sometimes a judgment on the benefit-risk relationship cannot be made and is then labeled "judgment cannot be made."

The benefit-risk judgment quality ρ_i of an intervention I_i, as recommended by GRADE, corresponds to the lowest meta-analysis quality of the critical preferences (step 4 of Algorithm 2). For example, given two preferences, "quality of life" and "somnolence" whose meta-analysis quality is "moderate" and, "low", respectively, the benefit-risk judgment quality is "low".

3.7. Evidence to decision: GRADE heuristics

Finally, the benefit-risk judgement and its quality for each intervention (I_i, j_i, ρ_i) are used to formulate the recommendation Rec_i for intervention I_i, which can be "in favor" or "against", depending on whether or not the benefit-risk relationship is favorable. The strength of the recommendation can be "strong" or "weak". Favorable recommendations are "strong" when the

Rule 1
IF acceptability preference favors placebo
$(e_{p_i} > 1.00$ where $kind\_of(p_i) = acceptability)$
THEN the benefit-risk judgment is
"probably favorable to the placebo"

Rule 3a
IF acceptability preference
$(e_{p_i} > 1.00$ where $kind_of(p_i) =$acceptability)
AND
no safety preference is deemed critical
$(\forall p_i$ such that $kind\_of(p_i) =$ safety then $p_i < 7)$
AND
the $e_{p_i} > 0.5$ on all the efficacy preferences
$(p_i$ such that $kind\_of(p_i) =$ efficacy)
THEN the benefit-risk judgment is
"favorable to the intervention"

Fig. 3. Examples of rules. e_{p_i} is the pooled effect of a meta-analysis outcome regarding a given preference p_i.

Table 4
GRADE heuristic from [30].

		benefit-risk judgment quality			
		high	moderate	low	very-low
benefit-risk judgment	favorable to the intervention	strong in favor	strong in favor	weak in favor	weak in favor
	probably favorable to the intervention	weak in favor	weak in favor	weak in favor	weak in favor
	neither favorable nor unfavorable	strong against	weak against	weak against	weak against
	probably favorable to the placebo	weak against	weak against	weak against	weak against
	favorable to the placebo	strong against	strong against	weak against	weak against

benefit-risk relationship is favorable and the quality of this relationship is high or moderate; otherwise, they are labeled "weak". Recommendations "against" the use of the intervention are "strong" when the benefit-risk relationship is neither favorable nor unfavorable and its certainty is "high", or when the relationship is unfavorable and its certainty is "high" or "moderate"; otherwise, they are labeled "weak". Table 4 summarizes the heuristic that we used to decide the recommendation. When the benefit-risk judgment is "judgment cannot be made", no recommendation is formulated.

As a result of this step, the user is provided with a list of interventions, with recommendations $(I_1, Rec_1), ...(I_S, Rec_S)$, sorted according to Rec_i (from "strongly in favor" to "strongly against"). This is personalized because interventions gathered from studies matching a patient's basal characteristics have been taken into account. Similarly, the recommended therapies are participatory because only RCTs that match the user's (patient and clinician) preferences are considered.

4. APPRAISE-RS implementation

The APPRAISE-RS methodology has been implemented to develop a recommender system for ADHD. ADHD is a prevalent psychiatric disorder that affects 5.9% of youths and 2.5% of adults [39]. It has a significant impact on society and CPGs recommendations lack effective implementation in everyday routine care [40]. Therefore, the need for a new therapy recommendation approach is of paramount importance because it will improve the treatment decision-making process.

The APPRAISE-RS implementation for ADHD has resulted in the APPRAISE-RS/TDApp recommender system (https://tdapp.org/). To support the patient's preference selection, the patient (or caregiver, depending on the patient's age) is informed about ADHD as a health condition and possible treatments.

The meta-analysis was implemented using R software (version 3.2.3) [41] through the meta and metafor libraries. The RBSs and the remaining elements of APPRAISE-RS/TDApp were implemented using Java.

4.1. Clinical studies

To carry out the meta-analysis, APPRAISE-RS/TDApp uses the information stored in the Minerva database (https://minervadatabase.org/en). The Minerva database contains comprehensive information on RCTs that have investigated the efficacy and safety of pharmacological interventions for ADHD. These RCTs are identified using systematic search techniques on Medline, Cochrane CENTRAL, Psycinfo, clinicaltrials.gov, clinicaltrialsregister.eu, and controlled-trials.com. Through a system of

weekly alerts, the contents of the Minerva database are updated each time that new studies are identified. At the time of the simulation study, the Minerva database had stored data from 348 RCTs (from 1987 to 2021) that include administrative information, study methods, patient characteristics, study results, and risk of bias. The usefulness of the Minerva database has been shown in previous studies [42].

Minerva, requires some preprocessing for the information to fit the structure required by APPRAISE-RS (Table 1). In this prototype version, APPRAISE-RS/TDApp considers data for 18 treatment goals or outcomes relevant to the treatment of patients with ADHD, including efficacy outcomes (i.e., ADHD symptom severity, clinical global impression, prevention of substance use disorder, drug use, accidents, academic performance, quality of life (QoL)), safety outcomes (i.e., appetite loss, insomnia, drowsiness, dizziness, dry mouth, tics, seizures, vomiting, syncope/faint, treatment discontinuation due to any adverse event), and acceptability outcomes (i.e., all-cause treatment discontinuation). For a detailed description of the final ADHD clinical studies database, see https://caleta.udg.edu/git/eXiT_Research_Group/TDAH_data.

4.2. Patient data

A case study was conducted using simulated data for 28 patients. The patients' demographics and clinical characteristics were simulated in a way that reflected the diversity of patients in a real-world setting and not the distribution of characteristics found in epidemiological studies of ADHD patients. A team of six members, including four psychiatrists and two EBM experts, set the demographic and clinical characteristics, and the patient's and clinician's preferences. In total, 20 of the patients were children or adolescents, and eight were adults. ADHD severity was moderate in 13 patients and severe in 15 patients. Meanwhile, 18 patients had a comorbidity disorder: oppositional defiant disorder (N = 4), dyslexia (N = 1), depressive disorder (N = 1), bipolar disorder (N = 1), autism spectrum disorder (N = 1), tics (N = 2), borderline personality disorder (N = 1), anorexia(N = 2), tobacco use disorder (N = 5), cannabis use disorder (N = 1), anorexia nervosa (N = 2), and epilepsy (N = 1). Six patients were receiving concomitant pharmacological treatment. Patients with comorbid conditions, other than tobacco use disorder, were considered to be complex. The full dataset of the simulated patients is available at https://caleta.udg.edu/git/eXiT_Research_Group/TDAH_data.

4.3. Experimental set-up

The APPRAISE-RS/TDApp recommendations were compared to those from recent and relevant clinical practice guidelines that provide advice on ADHD treatment, as follows: American Academy of Pediatrics (AAP) CPG [43], Canadian ADHD Resource Alliance (CADDRA) CPG [44], National Institute for Health and Care Excellence(NICE) CPG from the UK [45], and the Spanish CPG [46]. In this study, four clinicians provided the treatment that would have prescribed to the 28 simulated patients: two were child/adolescent psychiatrists and two were adult psychiatrists.

Moreover, after the clinicians' prescriptions were provided, the clinicians were requested to complete a satisfaction survey asking clinicians to rate from 0 (completely disagree) to 3 (completely agree) the following statement "I would feel comfortable prescribing this patient the treatment recommended by APPRAISE/TDApp".

5. Results

The results were analyzed according to the number of recommendations provided by APPRAISE-RS/TDApp, and to the similarities to the recommendations provided by the CPGs and the clinician's prescriptions.

5.1. Automated recommendations

Table 5 shows a summary of the number of recommendations provided by each recommender. On average, 1.96 recommendations per patient were made using APPRAISE-RS/TDApp. These recommendations were generated almost instantly upon entering clinical features and preferences into the system. The number of recommended interventions using APPRAISE-RS/TDApp was less than by using CPGs in almost all cases. This means that APPRAISE-RS/TDApp provides more specific recommendations than CPGs. Psychiatrists prescribe a single intervention, and are therefore not shown in the table.

The comparison between APPRAISE-RS/TDApp's recommendations and those of clinicians and CPGs was investigated in more depth by using a distance measure. The distance measure was defined within the [0,3] range, which takes into account the similarity among treatments according to the NbN ontology [47] (see https://caleta.udg.edu/git/eXiT_Research_Group/APPRAISE-intervention-distances for further details). The distance between the APPRAISE-RS/TDApp recommendations, the CPG recommendations, and the clinicians' prescription was moderate (Table 6). The distance between the interventions recommended by APPRAISE-RS/TDApp, and those recommended by the CPGs and the clinicians was modest, irrespective of the patient's age and the patient's complexity. This finding might mean that APPRAISE-RS/TDApp recommendations, while not identical, did not deviate widely from the CPG recommendations or the treatment prescribed by the psychiatrists in clinical practice. Meanwhile, the median of the distance values shown in Table 6 tells us that APPRAISE-RS/TDApp recommendations include at least one intervention that is identical or very similar to other drugs recommended by CPGs, or prescribed by psychiatrists. This means that APPRAISE-RS/TDApp provides acceptable recommendations in clinical practice.

Table 7 provides information about the interventions collected in the APPRAISE-RS/TDApp recommendations. The most frequently recommended intervention was high doses of methylphenidate, while the most frequent recommendation 'against' was

Table 5
Number of recommendations provided by APPRAISE-RS/TDApp in the 28 simulated patients, and comparison with the psychiatrists' prescriptions and the CPG recommendations.

	ALL	CHILDREN AND ADOLESCENTS	ADULTS	NONCOMPLEX PATIENTS	COMPLEX PATIENTS
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
APPRAISE-RS/TDApp	1.96 (1.50)	2.10 (1.68)	1.63 (0.92)	2.56 (1.46)	1.17 (1.19)
CPG USA	15.30 (2.36)	15.30 (2.36)	NA	16.00 (0.00)	14.44 (3.43)
CPG Canada	5.11 (1.34)	5.40 (1.31)	4.38 (1.19)	5.5 (1.03)	4.58 (1.56)
CPG Spain	6.64 (2.11)	7.50 (0.89)	4.50 (2.78)	6.25 (2.62)	7.17 (1.03)
CPG UK	2.29 (0.71)	2.00 (0.00)	3.00 (1.07)	2.38 (0.81)	2.17 (0.58)
All CPG	6.72 (4.81)	7.55 (5.11)	3.96 (1.90)	6.81 (4.92)	6.60 (4.72)

p-value <0.05 for all APPRAISE-RS/TDApp vs CPG comparisons in all simulated patients except for APPRAISE-RS/TDApp vs CPG UK comparison
p-value <0.05 for all APPRAISE-RS/TDApp vs CPG comparisons in children and adolescents except for APPRAISE-RS/TDApp vs CPG UK comparison
p-value <0.05 for all APPRAISE-RS/TDApp vs CPG comparisons in adults
p-value <0.05 for all APPRAISE-RS/TDApp vs CPG comparisons in noncomplex patients except for APPRAISE-RS/TDApp vs CPG UK comparison
p-value <0.05 for all APPRAISE-RS/TDApp vs CPG comparisons in complex patients

Table 6
Distance between the interventions recommended by APPRAISE-RS/TDApp and those by CPGs and clinicians.

	Mean (SD)	Median Min	Median Max
GPG USA	1.72 (0.70)	0.00	2.00
CPG Canada	1.24 (0.94)	0.00	2.00
CPG Spain	1.57 (0.80)	0.00	2.00
CPG UK	1.46 (0.99)	0.13	2.00
All CPG	1.48 (0.88)	0.00	2.00
Clinicians	1.34 (0.98)	0.33	2.00

Table 7
Detailed results of the last steps of APPRAISE-RS/TDApp when applied to 28 simulated ADHD patients. Interventions could be high or low dose.

	Risk-benefit judgment					Quality of the risk-benefit assessment			Recommendations					
	In favor	Probably in favor	Probably not in favor	Not in favor	Neither in favor nor against	High	Moderate	Low	Very low	Strong in favor	Weak in favor	Weak against	Strong against	Any recommendation
Patients														
Mean	0.5	1.4	0.5	0.6	0.0	0.0	0.0	0.6	2.3	0.0	2.0	1.0	0.0	2.9
SD	0.6	1.2	0.8	1.1	0.2	0.0	0.0	0.9	1.7	0.0	1.5	1.4	0.0	1.8
Median	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	2.0	0.0	0.0	2.0
Min	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Max	2.0	4.0	3.0	3.0	1.0	0.0	0.0	3.0	6.0	0.0	5.0	5.0	0.0	7.0
Interventions														
Atomoxetine high	17.9	32.1	7.1	17.9	0.0	0.0	0.0	10.7	64.3	0.0	50.0	25.0	0.0	75.0
Atomoxetine low	0.0	28.6	10.7	0.0	3.6	0.0	0.0	32.1	10.7	0.0	28.6	14.3	0.0	42.9
Dexmethylphenidate high	0.0	0.0	3.6	0.0	0.0	0.0	0.0	0.0	3.6	0.0	0.0	3.6	0.0	3.6
Guanfacine high	0.0	17.9	3.6	10.7	0.0	0.0	0.0	14.3	17.9	0.0	17.9	14.3	0.0	32.1
Guanfacine low	0.0	3.6	0.0	0.0	0.0	0.0	0.0	0.0	3.6	0.0	3.6	0.0	0.0	3.6
Lisdexamfetamine high	0.0	28.6	0.0	0.0	0.0	0.0	0.0	3.6	25.0	0.0	28.6	0.0	0.0	28.6
Lisdexamfetamine low	0.0	7.1	0.0	0.0	0.0	0.0	0.0	3.6	3.6	0.0	7.1	0.0	0.0	7.1
Methylphenidate high	32.1	21.4	3.6	3.6	0.0	0.0	0.0	0.0	60.7	0.0	53.6	7.1	0.0	60.7
Methylphenidate low	0.0	7.1	3.6	0.0	0.0	0.0	0.0	0.0	10.7	0.0	7.1	3.6	0.0	10.7
Modafinil high	0.0	0.0	0.0	21.4	0.0	0.0	0.0	0.0	21.4	0.0	0.0	21.4	0.0	21.4
Modafinil low	0.0	0.0	10.7	0.0	0.0	0.0	0.0	0.0	10.7	0.0	0.0	10.7	0.0	10.7

high dose atomoxetine. These results align with the usual pharmacological treatment in clinical practice. No non-approved medication for ADHD was recommended.

5.2. Updated recommendations

The ADHD database contains RCT publications from 1987 to 2021. Viloxazine, which was authorized in 2021 in the United States, is already considered in APPRAISE-RS/TDApp (see Table 2), even though it is still not in the AAP CPG (USA).

5.3. Personalized recommendations

There were 20 recommendations using APPRAISE-RS/TDApp: three using AAP CPG, seven using CADDRA CPG, and four using both the Spanish and NICE GPGs. CPGs have one recommendation that is frequently given to most patients, while APPRAISE-RS/TDApp has no single recommendation given to the majority of patients (i.e., 20 different recommendations for 28 patients). These findings support the idea that APPRAISE-RS/TDApp formulates more diverse recommendations than CPGs, which is likely to indicate a higher degree of personalization using APPRAISE-RS/TDApp than CPGs.

5.4. Participatory recommendations

To analyze the participatory dimension of the methodology, we carried out two additional experiments: the first only had the mandatory preference that clinician's often include (i.e., improve symptoms and avoid dropping treatment); and the second had random preferences. Symptom severity and acceptability preferences were always included. The results were compared with the previous setting labeled "baseline" and are shown in Table 8. Changing the preferences resulted in different APPRAISE-RS/TDApp recommendations. Table 8 shows that few interventions remained the same. This validates the impact of patient participation in APPRAISE-RS.

5.5. Clinicians' satisfaction

Table 9, shows the clinicians' satisfaction results (within the [0.3] range). The scores ranged from 0.3 to 2.7, with a median of 2.0. The clinicians' score was similar between child-adolescent and adult patients, as well as between noncomplex and complex patients. This finding is consistent with the previous finding, which shows that the distance between the interventions recommended by APPRAISE-RS/TDApp and the treatment prescribed by the psychiatrists was modest in all results.

6. Discussion

The results of this study using simulated data show that APPRAISE-RS/TDApp formulates valid, updated, personalized, and participatory recommendations. The following three findings support the validity of APPRAISE-RS/TDApp recommendations. First, the interventions most frequently recommended are underpinned by the largest body of evidence in the literature, and are amongst the most frequently recommended and used drugs in clinical practice. Second, clinician satisfaction was good. Third, the minimum distance between APPRAISE-RS/TDApp recommendations, CPG recommendations, and clinicians' prescriptions is 0 (or close to 0). This means that APPRAISE-RS/TDApp always includes at least one intervention that is identical, or very similar, to other drugs recommended by CPGs or prescribed by psychiatrists.

APPRAISE-RS/TDApp formulates updated recommendations because it uses the information stored in the Minerva database, which is updated through a system of weekly alerts. For this reason, APPRAISE-RS/TDApp analyses recently published RCTs and recently marketed drugs that are not considered by any CPG.

APPRAISE-RS/TDApp formulates personalized recommendations. This personalization arises from filtering out only those RCTs in which the patient could have participated, as well as considering the outcomes/preferences that are deemed relevant by the patient and the treating physician. The fact that the number of different recommendations is higher and the number of pharmacological interventions within each recommendation is lower with APPRAISE-RS/TDApp than with CPGs supports the idea that treatment recommendations are more personalized with APPRAISE than in CPGs.

APPRAISE-RS/TDApp's recommendations are participatory because the patients and the treating physicians select the relevant outcomes to be considered when analyzing the benefit-risk relationship of the interventions. Therefore, changing the preferences results in a different set of interventions being recommended. In this regard, some attention should be given to the use of the tool by both physicians and patients because selecting all preferences with the maximum value with the aim to achieve a Utopia (i.e., an efficacious intervention with no side effects) could lead to no system recommendation at all. This behavior would have unexpected outcomes because the system could not find an answer for a high constrained query. Therefore, further functionalities of the tool should include some control on the number and the preferences' score expressed by the users.

6.1. Limitations

The research presented in this work has observed some limitations. First, APPRAISE-RS, as any other recommender system, suffers from the cold start problem [48] (i.e., the knowledge required to start providing useful recommendations). In particular, APPRAISE-RS assumes the availability of a database of clinical studies on the disease that the analysis is focused on. This may be an issue given that not all diseases have a robust literature available, which is for our case study on ADHD. Meanwhile, when there is a corpus of literature available, current text mining techniques are limited regarding the automatic generation of a database with an

Table 8
Number of recommendations with different preference settings and the same patient basal data. Removed: number of recommendations present in the baseline setting that are no longer present in the new experiment. Added: number of recommendations that are provided in the new experiments recommendations but were not in the baseline setting.

		P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	P22	P23	P24	P25	P26	P27	P28	Average	SD
Baseline	Number	2	2	2	1	1	1	2	1	1	2	1	2	2	2	2	7	2	0	2	4	3	6	6	4	7	6	4	1	2.714	1.997
Often	Number	4	5	4	3	1	1	2	0	1	2	0	3	1	3	5	4	1	0	2	3	1	2	2	0	2	2	0	1	1.964	1.503
	Removed	0	0	1	0	1	0	0	1	0	0	1	2	2	2	2	3	1	0	0	1	2	4	4	4	5	4	4	0	1.571	1.620
	New	2	3	3	2	1	0	0	0	0	0	0	3	1	3	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0.821	1.389
Random	Number	0	1	0	0	2	0	1	0	0	0	2	1	1	4	0	1	0	0	1	1	1	4	4	1	10	0	1	0	1.286	2.106
	Removed	2	2	2	1	1	1	1	1	1	2	1	2	2	0	2	6	2	0	1	3	2	5	5	3	5	6	3	1	2.250	1.691
	New	0	1	0	0	2	0	0	0	0	0	2	1	1	2	0	0	0	0	0	0	0	3	3	0	8	0	0	0	0.821	1.701

Table 9
Clinician satisfaction with APPRAISE-RS/TDApp recommendations.

	Mean (SD)	Min	Max
All	2.15 (0.94)	0.00	3.00
Children and adolescents	2.13 (0.98)	0.00	3.00
Adults	2.20 (0.83)	0.00	3.00
NonComplex patients	2.47 (0.76)	0.00	3.00
Complex patients	1.74 (1.01)	0.00	3.00

accurate result for real clinical practice [32]. We expect that this limitation will be overcome soon thanks to the recent advances of DL in extracting information from medical publications databases [49,50].

The experiments that we carried out in this study are based on simulated data. APPRAISE-RS/TDApp is considered to be a Medical Device according to the EU regulation [51]. Because APPRAISE-RS/TDApp aims to identify which treatment is more likely to be efficacious and safe for a given patient, and thus is expected to influence treatment decision making, preclinical studies are an ethical mandate before performing a clinical study. In the context of AI-assisted decision-making, studies using simulated data seem to be the best option and have been shown useful in the development of clinical decision support systems [52]. Other possibilities for testing the system include using retrospective data (historical patient data) [15], or prospective data in the context of an RCT. Although the former approach is not possible due to the fact that we have neither clinical nor patient preferences stored in the current healthcare systems, the latter approach is underway (https://clinicaltrials.gov/ct2/show/NCT04228094), although the current context of the pandemic is delaying its execution. Once the clinical trial has been conducted, the next development step should be the integration of the application with the Hospital Information System, to automatically gather the basal information from patients, and register back the treatment chosen by the clinician.

Finally, to date, APPRAISE-RS/TDApp is limited to pharmacological interventions. Future updates of this tool should also include other commonly used interventions, such as psychotherapy.

6.2. Academic and clinical implications

At the point of care, having a tool that formulates participatory treatment recommendations will enable patients to have an active role in the decision-making process, which is likely to improve treatment adherence, and consequently the clinical outcomes. Moreover, personalization supported by a computer tool can help to harmonize the treatment prescribed by different clinicians to the same or similar patients. Nevertheless, it must be stressed that the APPRAISE-RS recommendation should help and not replace clinical judgement.

APPRAISE opens several lines of research for the academic community. First, considering indirectness and similarities, as discussed earlier. A more interesting challenge is to improve the GRADE heuristics by considering all of the patient's and clinician's preferences, irrespective of their score. Currently, following the GRADE heuristics, only those preferences scoring 7 or above are analyzed and the remaining preferences are ruled out, which leads to a loss of information. Multi-criteria decision methods that are available from AI could handle all preferences and weight them appropriately. Moreover, explainability issues can be easily incorporated in APPRAISE-RS because all of the evidence generated throughout the recommendation-making process can be made available to the clinicians. This is relevant because explainable capabilities induce trust in doctors [53]. More challenging research proposals include the adaptation of the system according to its own experience.

7. Conclusions

CPGs enable evidence-based medicine to be put into practice, and some AI-based tools facilitate keeping CPGs up-to-date. Meanwhile, other AI tools support tailoring recommendations to the patient's characteristics (personalization), or provide health condition information (as a kind of user participation). This research presents APPRAISE-RS, which is a methodology that addresses CPG development and maintenance in an holistic way. Consequently, it generates treatment recommendations from medical publications (automation); keeps the knowledge of a given disease updated; applies the GRADE methodology according to the patient's characteristics (personalization); and takes the end-users, clinicians, and patients into account (participatory).

APPRAISE-RS was deployed for ADHD, resulting in APPRAISE-RS/TDApp, which has been experimentally tested in 28 simulated patients. The results show the differences between state-of-the art CPGs and APPRAISE-RS regarding the specificity of the recommendations (i.e., fewer number of drugs and more variety of recommendations), and support the personalization and participatory dimensions of APPRAISE-RS. Clinician satisfaction validates the usefulness of APPRAISE-RS/TDApp's recommendations provided by APPRAISE-RS. Our current research involves testing APPRAISE-RS in a real setting, but this progress has been hindered by the COVID-19 pandemic.

Further research could consider the meta-analysis step of the GRADE method to include studies that could be triggered due to similarities with the patient's characteristics, instead of being an exact match. Moreover, the use of multi-criteria decision-making techniques to weight the end-user's preferences according to the scores provided by the patient is an interesting line of research that could be explored in the future.

Funding statement

This work was supported by European Regional Development Fund (ERDF), the Spanish Ministry of the Economy, Industry and Competitiveness (MINECO) and the Carlos III Research Institute [PI19/00375], Fundació Pascual i Prats & Campus Salut, UdG [AIN2018E], Generalitat de Catalunya [2017 SGR 1551].

Author contribution statement

Beatriz López: Conceived and designed the experiments; Analyzed and interpreted the data; Wrote the paper. **Oscar Raya:** Performed the experiments; Analyzed and interpreted the data. **Evgenia Baykova, Marc Saez, David Rigau, Ruth Cunill, Sacramento Mayoral, Carme Carrion, Domènec Serrano:** Contributed reagents, materials, analysis tools or data. **Xavier Castells:** conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgements

We would like to thank Marc González for his support in the initial implementation of some components of the prototype.

References

[1] B. Djulbegovic, G.H. Guyatt, Progress in evidence-based medicine: a quarter century on, Lancet (London, England) 390 (10092) (2017) 415–423, https://doi.org/10.1016/S0140-6736(16)31592-6.
[2] P. Alonso-Coello, A. Oxman, et al., GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 2: clinical practice guidelines, BMJ (Clinical research ed.) 353 (2016) i2089, https://doi.org/10.1136/BMJ.I2089.
[3] V. Zamborlini, Q. Hu, et al., Knowledge-driven paper retrieval to support updating of clinical guidelines, in: Knowledge Representation for Health Care, Springer International Publishing, Cham, 2017, pp. 71–89.
[4] Z. Huang, Q. Hu, et al., Identifying evidence quality for updating evidence-based medical guidelines, in: Knowledge Representation for Health Care, Springer International Publishing, Cham, 2015, pp. 51–64.
[5] E. Clark, E.F. Donovan, P. Schoettker, From outdated to updated, keeping clinical guidelines valid, Int. J. Quality Health Care 18 (3) (2006) 165–166, https://doi.org/10.1093/INTQHC/MZL007.
[6] N. Steel, A. Abdelhamid, et al., A review of clinical practice guidelines found that they were often based on evidence of uncertain relevance to primary care patients, J. Clin. Epidemiol. 67 (11) (2014) 1251–1257, https://doi.org/10.1016/J.JCLINEPI.2014.05.020.
[7] B. Austad, I. Hetlevik, B. Mjølstad, A. Helvik, Applying clinical guidelines in general practice: a qualitative study of potential complications, BMC Family Practice 17 (1) (2016), https://doi.org/10.1186/S12875-016-0490-3.
[8] M. Armstrong, J. Bloom, Patient involvement in guidelines is poor five years after institute of medicine standards: review of guideline methodologies, Res. Involv. Engag. 3 (1) (2017), https://doi.org/10.1186/S40900-017-0070-2.
[9] D. Arts, A. Voncken, S. Medlock, et al., Reasons for intentional guideline non-adherence: a systematic review, Int. J. Med. Inform. 89 (2016) 55–62, https://doi.org/10.1016/J.IJMEDINF.2016.02.009.
[10] V.C. Correa, L.H. Lugo-Agudelo, et al., Individual, health system, and contextual barriers and facilitators for the implementation of clinical practice guidelines: a systematic metareview, Health Res. Policy Syst. 18 (1) (2020) 1–11, https://doi.org/10.1186/S12961-020-00588-8.
[11] G. Bosse, J.P. Breuer, C. Spies, The resistance to changing guidelines: what are the challenges and how to meet them, Best Practice Res. Clin. Anaesthesiol. 20 (3) (2006) 379–395, https://doi.org/10.1016/J.BPA.2006.02.005.
[12] C. Chen, K. Chen, C.-Y. Hsu, W.-T. Chiu, Y.-C.J. Li, A guideline-based decision support for pharmacological treatment can improve the quality of hyperlipidemia management, Comput. Methods Programs Biomed. 97 (3) (2010) 280–285, https://doi.org/10.1016/j.cmpb.2009.12.004.
[13] H. Kondylakis, A. Bucur, et al., Patient empowerment for cancer patients through a novel ICT infrastructure, J. Biomed. Inform. 101 (2020) 103342, https://doi.org/10.1016/j.jbi.2019.103342.
[14] M. Peleg, Computer-interpretable clinical guidelines: a methodological review, J. Biomed. Inform. 46 (2013) 744–763, https://doi.org/10.1016/J.JBI.2013.06.009.
[15] B. López, F. Torrent-Fontbona, et al., HTE 3.0: knowledge-based systems in cascade for familial hypercholesterolemia detection and dyslipidemia treatment, Expert Syst. (2021), https://doi.org/10.1111/exsy.12835.
[16] L. Verboven, T. Calders, et al., A treatment recommender clinical decision support system for personalized medicine: method development and proof-of-concept for drug resistant tuberculosis, BMC Med. Inform. Decis. Mak. 22 (1) (2022) 1–11, https://doi.org/10.1186/s12911-022-01790-0.
[17] M. Eghbali-Zarch, R. Tavakkoli-Moghaddam, et al., Pharmacological therapy selection of type 2 diabetes based on the SWARA and modified MULTIMOORA methods under a fuzzy environment, Artif. Intell. Med. 87 (2018) 20–33, https://doi.org/10.1016/J.ARTMED.2018.03.003.
[18] W. Van Woensel, S.S.R. Abidi, S.R. Abidi, Decision support for comorbid conditions via execution-time integration of clinical guidelines using transaction-based semantics and temporal planning, Artif. Intell. Med. 118 (2021) 102127, https://doi.org/10.1016/J.ARTMED.2021.102127.
[19] V. Zamborlini, R. van der Heijden, A. ten Teije, Filtering clinical guideline interactions with pre-conditions: a case study on diabetes guideline, CEUR Workshop Proc. 2237 (2018) 26–33.
[20] X. Liu, C.H. Chen, et al., A DNA-based intelligent expert system for personalised skin-health recommendations, IEEE J. Biomed. Health Inform. 24 (11) (2020) 3276–3284, https://doi.org/10.1109/JBHI.2020.2978667.
[21] S. Garg, Drug recommendation system based on sentiment analysis of drug reviews using machine learning, in: Proceedings of the Confluence 2021: 11th International Conference on Cloud Computing, Data Science and Engineering, 2021, pp. 175–181, arXiv:2104.01113.
[22] F. Gräßer, S. Beckert, et al., Therapy decision support based on recommender system methods, J. Healthcare Eng. (2017), https://doi.org/10.1155/2017/8659460.
[23] C.I. Ossai, N. Wickramasinghe, Intelligent therapeutic decision support for 30 days readmission of diabetic patients with different comorbidities, J. Biomed. Inform. 107 (2020) 103486, https://doi.org/10.1016/j.jbi.2020.103486.

[24] A.P. Wright, A.T. Wright, et al., The use of sequential pattern mining to predict next prescribed medications, J. Biomed. Inform. 53 (2015) 73–80, https://doi.org/10.1016/J.JBI.2014.09.003.
[25] Ö. Taçyıldız, D. Çelik Ertuğrul, A decision support system on the obesity management and consultation during childhood and adolescence using ontology and semantic rules, J. Biomed. Inform. 110 (2020) 103554, https://doi.org/10.1016/j.jbi.2020.103554.
[26] F. Torrent-Fontbona, B. Lopez, Personalized adaptive CBR bolus recommender system for Type 1 diabetes, IEEE J. Biomed. Health Inform. 23 (1) (2019) 387–394, https://doi.org/10.1109/JBHI.2018.2813424.
[27] B. Rose-Davis, W. Van Woensel, et al., Semantic knowledge modeling and evaluation of argument theory to develop dialogue based patient education systems for chronic disease self-management, Int. J. Med. Inform. 160 (2022) 104693, https://doi.org/10.1016/J.IJMEDINF.2022.104693.
[28] S. Wilk, M. Michalowski, W. Michalowski, D. Rosu, M. Carrier, M. Kezadri-Hamiaz, Comprehensive mitigation framework for concurrent application of multiple clinical practice guidelines, J. Biomed. Inform. 66 (2017) 52–71, https://doi.org/10.1016/J.JBI.2016.12.002.
[29] J. Fdez-Olivares, E. Onaindia, L. Castillo, J. Jordán, J. Cózar, Personalized conciliation of clinical guidelines for comorbid patients through multi-agent planning, Artif. Intell. Med. 96 (2019) 167–186, https://doi.org/10.1016/J.ARTMED.2018.11.003.
[30] G. Guyatt, A.D. Oxman, et al., GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables, J. Clin. Epidemiol. 64 (4) (2011) 383–394, https://doi.org/10.1016/J.JCLINEPI.2010.04.026.
[31] B.R. Da Costa, E. Nüesch, et al., Combining follow-up and change data is valid in meta-analyses of continuous outcomes: a meta-epidemiological study, J. Clin. Epidemiol. 66 (8) (2013) 847–855, https://doi.org/10.1016/J.JCLINEPI.2013.03.009.
[32] B. Percha, Modern clinical text mining: a guide and review, Annu. Rev. Biomed. Data Sci. 4 (1) (2021) 165–187, https://doi.org/10.1146/ANNUREV-BIODATASCI-030421-030931.
[33] B. Derrick, P. White, Comparing two samples from an individual Likert question, Int. J. Math. Stat. 18 (3) (2017).
[34] M. Riera, X. Castells, et al., Discontinuation of pharmacological treatment of children and adolescents with attention deficit hyperactivity disorder: meta-analysis of 63 studies enrolling 11,788 patients, Psychopharmacology 234 (17) (2017) 2657–2671, https://doi.org/10.1007/S00213-017-4662-1.
[35] M.W. Cheung, R. Vijayakumar, A guide to conducting a meta-analysis, Neuropsychol. Rev. 26 (2) (2016) 121–128, https://doi.org/10.1007/S11065-016-9319-Z.
[36] J.M. Reps, J.M. Garibaldi, U. Aickelin, D. Soria, J.E. Gibson, R.B. Hubbard, A novel semisupervised algorithm for rare prescription side effect discovery, IEEE J. Biomed. Health Inform. 18 (2) (2014) 537–547, https://doi.org/10.1109/JBHI.2013.2281505.
[37] J. Deeks, J. Higgins, D. Altman, Analysing data and undertaking meta-analyses, in: J.P.T. Higgins, et al. (Eds.), Cochrane Handbook for Systematic Reviews of Interventions, Chapter 10, version 6.2 (updated February 2021), 2021, https://training.cochrane.org/handbook/current.
[38] C.H. Lee, S. Cook, et al., Comparison of two meta-analysis methods: inverse-variance-weighted average and weighted sum of Z-scores, Genomics Inform. 14 (4) (2016) 173–180, https://doi.org/10.5808/GI.2016.14.4.173.
[39] S.V. Faraone, T. Banaschewski, et al., The World Federation of ADHD International Consensus Statement: 208 Evidence-based conclusions about the disorder, https://doi.org/10.1016/j.neubiorev.2021.01.022, 2021.
[40] B. Libutzki, M. May, et al., Disease burden and direct medical costs of incident adult ADHD: a retrospective longitudinal analysis based on German statutory health insurance claims data, Eur. Psychiatr. 63 (1) (2020), https://doi.org/10.1192/j.eurpsy.2020.84.
[41] R Core Team R, A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2018, https://www.R-project.org/.
[42] X. Castells, M. Saez, et al., Placebo response and its predictors in Attention Deficit Hyperactivity Disorder: a meta-analysis and comparison of meta-regression and MetaForest, Int. J. Neuropsychopharmacol. (Aug 2021), https://doi.org/10.1093/IJNP/PYAB054.
[43] M.L. Wolraich, J.F. Hagan, et al., Clinical practice guideline for the diagnosis, evaluation, and treatment of attention-deficit/hyperactivity disorder in children and adolescents, Pediatrics 144 (4) (2019), https://doi.org/10.1542/PEDS.2019-2528.
[44] Canadian ADHD Resource Alliance, Canadian ADHD Practice Guidelines, Tech. Rep., CADDRA, Toronto ON, Canada, 2018.
[45] National Guideline Centre UK, Attention deficit hyperactivity disorder: Diagnosis and management, Tech. Rep., National Institute for Health and Care Excellence (UK), London, 2018.
[46] Grupo de trabajo de la Guía de Práctica Clínica sobre las Intervenciones Terapéuticas en el Trastorno por Déficit de Atención con Hiperactividad (TDAH), Guía de Práctica Clínica sobre las Intervenciones Terapéuticas en el TDAH., Tech. rep., Ministerio de Sanidad, Servicios Sociales e Igualdad, Instituto Aragonés de Ciencias de la Salud (IACS); Guías de Práctica Clínica en el SNS (2017).
[47] J. Zohar, S. Stahl, et al., A review of the current nomenclature for psychotropic agents and an introduction to the neuroscience-based nomenclature, Eur. Neuropsychopharmacol. 25 (12) (2015) 2318–2325, https://doi.org/10.1016/j.euroneuro.2015.08.019.
[48] B. Lika, K. Kolomvatsos, S. Hadjiefthymiades, Facing the cold start problem in recommender systems, Expert Syst. Appl. 41 (4) (2014) 2065–2073, https://doi.org/10.1016/J.ESWA.2013.09.005.
[49] J. Chen, X. Sun, X. Jin, R. Sutcliffe, Extracting drug-drug interactions from no-blinding texts using key semantic sentences and ghm loss, J. Biomed. Inform. 9 (2022), https://doi.org/10.1016/J.JBI.2022.104192.
[50] W. Kim, L. Yeganova, D.C. Comeau, W.J. Wilbur, Z. Lu, Towards a unified search: improving pubmed retrieval with full text, J. Biomed. Inform. (2022), https://doi.org/10.1016/J.JBI.2022.104211.
[51] European Commission, Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EE, Off. J. Eur. Union 117 (5.5.2017) (2017) 1–175.
[52] V. Sintchenko, E. Coiera, J.R. Iredell, G.L. Gilbert, Comparative impact of guidelines, clinical data, and decision support on prescribing decisions: an interactive web experiment with simulated cases, J. Am. Med. Inform. Assoc. 11 (1) (2004) 71–77, https://doi.org/10.1197/JAMIA.M1166.
[53] E. Contempré, Z. Szlávik, E. Velazquez-Godinez, A. ten Teije, I. Tiddi, Towards explained treatment search results: feature analysis and explanation formulation, in: First International Workshop on eXplainable AI in Healthcare (AIME), 2021.