Assessing the robustness of clinical trials by estimating Jadad's score using artificial intelligence approaches

Tiphaine Casy, Alexis Grasseau, A Charras, Bénédicte Rouvière, Jacques-Olivier Pers, Nathan Foulquier, Alain Saraux

Assessing the robustness of clinical trials by estimating Jadad's score using artificial intelligence approaches

Tiphaine Casy (1), Alexis Grasseau (2), Amandine Charras (3) Bénédicte Rouvière (4)(5) Jacques-Olivier Pers (5), Nathan Foulquier *(5) & Alain Saraux *(5)(6)

Affiliation

(1) LTSI MediCIS team, UMR 1099, University of Rennes 1, Inserm, Rennes, France
(2) MICMAC, UMR 1236, University of Rennes 1, Inserm, Rennes, France
(3) Department of Women's & Children's Health, Institute of Life Course and Medical Sciences, University of Liverpool, Liverpool, UK
(4) Department of internal medicine and pneumology, La Cavale Blanche, CHRU, Brest, France.
(5) LBAI, UMR 1227, University of Brest, Inserm, Brest, France
(6) Rheumatology Unit, Centre National de Référence des Maladies Auto-Immunes Rares (CERAINO), CHRU, Brest, France

* Nathan Foulquier and Alain Saraux are co-last authors

Corresponding author:

Prof. Alain Saraux

Rheumatology Unit, Hôpital de la Cavale Blanche, BP 824, F 29609 Brest cedex, France

E-mail: alain.saraux@chu-brest.fr

Highlight

- We built a program based on artificial intelligence approaches to assess the robustness of a clinical trial using the Jadad score.
- The program consists of five Recursive Neural Networks (RNN), each of which is trained to identify a specific item constituting the Jadad scale.
- After training, the algorithm achieved excellent accuracy on two separate validation sets

Abstract

Background: Clinical trials are essential in medical science and are currently the most robust strategy for evaluating the effectiveness of a treatment. However, some of these studies are less reliable than others due to flaws in their design. Assessing the robustness of a clinical trial can be a very complex and time-consuming task, with factors such as randomization, masking and the description of withdrawals needing to be considered.

Method: We built a program based on artificial intelligence (AI) approaches, designed to assess the robustness of a clinical trial by estimating its Jadad's score. The program is composed of five Recursive Neural Networks (RNN), each of them trained to spot one specific item constituting the Jadad's scale. After training, the algorithm was tested on two different validation sets (one from the original database: 35% of this database was used for validation and 65 % for training; one composed of 10 articles, out of the original database, for which the Jadad's score has been computed by each contributor of this study).

Result: After training, the algorithm achieved a mean accuracy of 96,2% (ranging from 93% to 98%) and a mean area under the curve (AUC) of 96% (ranging from 95% to 97%) on the first validation dataset. These results indicate good feature detection capacity for each of the five RNN. On the second validation dataset the algorithm extracted 100% of the item to retrieve for 70% of the articles and between 66% and 75% for 30% of the articles. Overall 85% of the items present in the second validation dataset were correctly extracted. None of the extracted items was misclassified.

Conclusion: We developed a program that can automatically estimate the Jadad's score of a clinical trial with a good accuracy. Automating the assessment of this metric could be very useful in a systematic review of the literature and will probably save clinicians time.

Key words: Clinical trials; artificial intelligence; robustness; systematic literature review
Funding: This research did not receive any specific grants from funding agent in the public, commercial and not-for-profit sectors

1. Introduction

In medicine, clinical trials are essential to study the effect of treatments on specific pathologies, to deepen and decipher the understanding of biological mechanisms (1–4). Thus, clinical trials designs have to be rigorous, scientifically robust and respectful of pre-established ethical rules (5). However, some of these studies are less reliable than others, due to flaws in their methodology (3). These shortcomings are usually the absence or inadequacy of randomization techniques and inappropriate or non-existent masking of patients. These considerations have given rise to the need to assess the quality of clinical trials (3,4,6).

The main biases in clinical trials are inadequate methodology and poor reporting, which have an impact on reproducibility of studies (1,7,8). In order to reduce bias, it is proposed that more reporting guidelines be adopted to improve study transparency and methodology (8). A paper from Vinkers et al. shows an improvement in this domain due to the last two points combined with increased awareness and mandatory registration of trials(1). However, low-impact factors journals remain the ones with the most identified biais and require special attention (1,8).

The first quality scale for clinical trials was developed in 1981, followed by 24 others over the next 5 years (9). These scoring systems consist of a list of various items measured usually by a binary score based on their presence or absence in the article/clinical trial report (1,9). To this date, lots of criteria can be considered to assess the quality of clinical trials : randomization, masking, allocation concealment, handling of withdrawals and dropouts, measures of variability, pre-specified analyses, stopping rules, statistical methods, baseline data, multiple addresses (1,6,9). Among them, criteria determined as particularly important to assess if the quality of clinical trials are related to randomization and blinding process (6,10). Additionally, description of study withdrawals and their reasons is also a key component to assess the quality of clinical trials (11).

Jadad's scale is one of the most widely used scores because of its simplicity and coverage of these three criteria. This scale represents the estimation of the robustness of a clinical trial by a numerical value (3,12). The Jadad's scale is also one of the few scoring systems designed using a standard scale development technique (13).

Computation of scores is still performed manually by human raters, which is a time-consuming process and may introduce risks of experimenter bias (13–17). Inter-rater

agreement can vary considerably depending on the item being assessed (13,14,16,17) or the purpose of the trial (12). These results highlight the need to reduce human bias when assessing the quality of a clinical trial. This task belongs to the field of natural language processing (NLP) , which has been solved in a few decades with a recursive neural network (RNN) (18,19). This type of deep learning model successfully establishes relationships between entities in a directed acyclic graph which can be applied to predict the meaning of sentences. To do this,, sentence's words and then the whole sentence are vectorized and represented as a dependency tree (19,20). One of the main advantages of RNNs, which improves the results in the NLP domain, is their ability to handle information patterns of different sizes (18,20). Therefore, the number of clinical trial publications continues to increase each year (from 2,119 in 2000 to 362,524 in 2020, according to clinicaltrials.gov). Thus, the need to be able to quickly assess their quality rises.

The aim of this study was to build a program based on AI approaches, designed to assess the robustness of a clinical trial through a simplified Jadad's score and then to validate the algorithm on a validation dataset.

2. Material & Methods

We designed an automated approach for quality assessment of randomized clinical trials with the Jadad scale *via* the use of AI techniques suitable for Natural Language Processing (NLP) tasks (figure 1).

2.1 The Jadad scale

We used a simplified version of the Jadad scale (15). This score is defined on a scale of 0 to 5 and consists of 5 items to be checked, the presence or absence of these items is associated with a score (table 1). A high value on the Jadad's scale characterizes a robust clinical trial (randomized, double-blind, with a correct explanation of the techniques used to perform randomization and blinding and a description of the study withdrawals). Each of the items on the Jadad's scale corresponds to the presence or absence of specific information in the text of an article describing a clinical trial.

In the original Jadad score, questions 4 and 5 are different from those presented in table 1. In the original score, the last two questions allow a response variability between -1, 0 and 1 regarding the compliance of described methodologies for randomization and masking with the study settings. We used a simplified version of the last two questions, considering only the presence or absence of a description of the randomization and/or masking process. This simplified version allows us to binarize the output of these two questions and, in doing so, to facilitate the training of the algorithms while maintaining a good estimation of the original Jadad score.

2.2 Recursive neural network

In order to assess the presence or absence of each of the items composing the Jadad scale, we used recursive neural networks (RNN). RNN are a type of artificial neural networks (ANN) dedicated to sequential information processing (21), specific cases of information processing where the context itself can be information. This type of architecture is therefore well suited to natural language processing (NLP). For each of the five items composing the Jadad's scale, we train a RNN to detect whether or not the information described by the corresponding item is present in a given sentence. RNN being a machine learning technique, we created a database of sentences extracted from abstract of published clinical trials to train these algorithms.

2.3 Database creation and constitution

We created a database of sentences using publications found on the Pubmed database[11]. We search articles describing clinical trials on various topics. We automatically assemble the title, abstract and methods section (when available) of each retrieved article using python scripts. The assembled text was then divided into sentences. For each sentence, two operators manually assigned the presence or absence of each of the 5 items constituting the Jadad's scale. In order to create the training database, we selected 1698 sentences from 617 articles. These sentences are distributed as follows: 493 for question 1; 154 for question 2; 111 for question 3; 157 for question 4 and 98 for question 5. 967 sentences do not contain any of the

[1] https://pubmed-ncbi-nlm-nih-gov.liverpool.idm.oclc.org/

5 items. Among these sentences, some are negations, such as “Third, this study was not double-blinded.”. Therefore, the database contains numerous possible situations including sentences with two or three different items.

2.4 Validation sets

We used two different validation sets:

- The first one came from the validation split of the original database: 35% of this database was used for validation and 65 % for training.
- The second validation dataset was composed of 10 articles (represented by their title and abstract) for which the Jadad’s score has been computed by each contributor of this study.

This dataset was designed as a concrete application case for our program in order to provide an additional proof of robustness in real situations of quality assessment.

2.5 Training and architecture of the RNNs

Training of the RNNs was performed on a modern laptop (8GB RAM, intel core i5). Algorithms were trained with a maximum number of epochs set to 350 and early stopping set to wait for a minimum of 3 epochs. All RNNs had the same architecture: one embedding layer, one long short-term memory layer and three dense layers. The activation function for the output layer was set to rectified linear.

2.6 Statistics

Performance of each of the five RNNs making up our program was measured using their accuracy and their AUC curve on the first validation dataset. AUC values provide a strong indicator of performance due to its robustness to overfitting. Each of the five RNN used a binary cross entropy loss function during training. Global performance of our assembled program was evaluated by computing accuracy of the score prediction on the second validation dataset, which was manually designed to be well balanced in terms of items to predict. All programs were written in python python 3.8.5 and RNN were built and trained using the keras 2.4.3 library (with tensorflow 2.3.1 backend).

3. Results

3.1 Training and first validation

RNN training stopped after an average of 23 epochs thanks to early stopping (figure 2). The RNN dedicated to item 1 detection reaches 20 epochs, 25 for item 2 detection, 37 for item 3, 13 for item 4 and 22 epochs for item 5. Each RNN was then tested on the first validation set and showed performances ranging from 93% to 98% accuracy with an AUC ranking from 95% to 97% (table 2). The loss and accuracy curves presented in figure 2 show very good performance and the absence of over-fitting.

3.2 Results for the second validation set

Of the 20 items to be retrieved from the 10 articles of the second validation dataset, 17 (85%) were correctly extracted by our program. Jadad's score was predicted with 100% accuracy for 7 articles (70 %), and between 66% and 75% accuracy for 3 articles (30%). No false positives were detected for any of the items. Results of the Jadad's score estimations and details of the extractions performed are presented in table 3. The lowest accuracy results always concern the fourth item, which deals with the details of the randomisation. The description of this last point in the sentences is heterogeneous, with no basic scheme, and item 4 is represented at 9.25% in the overall dataset. To reduce the poor detection of this item, the dataset could be incremented with more different sentences responding to item 4. Moreover, with the good results obtained on two distinct validation sets, we can be confident in the reliability of the Jadad's score approximation made by our program.

4. Discussion

We built a tool that automatically estimates the Jadad's score of a clinical trial by reading the text of the associated paper with five recursive neural networks, each of them trained to detect one of the five items constituting the Jadad score.

Our algorithm shows excellent performance in estimating the Jadad score of a clinical trial on a validation set and could help clinicians to quickly assess the quality of a paper, especially in the context of a systematic literature review (SLR). The Jadad's score is a widely used metric

to assess the quality of a clinical trial (32) and can therefore be used as a quality filter in SLR and meta analysis. Its simplicity enables it to be applied to a large number of topics since evaluated items are kept simple and are not specific to a particular subject. Among these items, the most important criteria for assessing the quality of a clinical trial are present (randomization and masking) (6,10). An estimation of Jadad's score is therefore a simple but very informative data point to assess the quality of a clinical trial.

AI approaches have been a game changer in the field of NLP (19). Their use in complex but structured contexts such as biomedical scientific publishing has led to applications that can save time for clinicians by automating time consuming tasks such as SLR (33,34). In this study we used AI approaches to perform specific information extraction from scientific publication to assess the quality of clinical trials by computing an estimation of their Jadad's score, and in doing so, further reduce the time clinicians need to spend analyzing articles.

The strength of this study lies in the size of the dataset used to train our programm (more than 600 articles manually ranked by multiple raters, resulting in thousands of sentences in the training database), the strong performances obtained on two separate validation datasets and the modularity of the architecture of the program itself : the parallelization of the execution of each RNN allows the addition of supplementary RNN dedicated to the detection of additional items. The imbalance in the training dataset was addressed by using the AUC metrics as a loss function and was overcome in the second validation dataset by selecting specific articles that contained all the items to be detected.

A limitation of this study is the difficulty to compute the exact Jadad score of the targeted clinical trial, as the relevant information for the score computation is not always contained in the abstract and title of the paper and the methods section was not always freely available. Additionally item 4 and 5 of the Jadad's scale have been simplified in order to provide a binary output, therefore, our tool only provides an estimation of the score.

In conclusion, our algorithm is able to automatically assess the quality of a clinical trial by computing an accurate estimation of its Jadad score using AI approaches. In the near future, we intend to use the same approach to develop estimators of other quality assessment scores and to use these algorithms to automate time-consuming aspects of literature meta- analysis such as ranking and filtering of clinical trials.

We expect that the need for such tools will increase in the near future due to the ever increasing number of clinical trials published each year.

Bibliography

1. Vinkers CH, Lamberink HJ, Tijdink JK, Heus P, Bouter L, Glasziou P, Moher D, Damen JA, Hooft L, Otte WM. The methodological quality of 176,620 randomized controlled trials published between 1966 and 2018 reveals a positive trend but also an urgent need for improvement. *PLoS Biol* (2021) **19**:e3001162. doi: 10.1371/journal.pbio.3001162
2. García-Perdomo HA, Díaz-Hung AM, Mejía LM. [Risk of bias assessment of clinical trials published in iberoamerican urological journals]. *Arch Esp Urol* (2015) **68**:615–626.
3. Silva Filho CR da, Saconato H, Conterno LO, Marques I, Atallah AN. [Assessment of clinical trial quality and its impact on meta-analyses]. *Rev Saude Publica* (2005) **39**:865–873. doi: 10.1590/s0034-89102005000600001
4. Jüni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. *BMJ* (2001) **323**:42–46. doi: 10.1136/bmj.323.7303.42
5. Umscheid CA, Margolis DJ, Grossman CE. Key concepts of clinical trials: a narrative review. *Postgrad Med* (2011) **123**:194–204. doi: 10.3810/pgm.2011.09.2475
6. Berger VW, Alperson SY. A general framework for the evaluation of clinical trial quality. *Rev Recent Clin Trials* (2009) **4**:79–88. doi: 10.2174/157488709788186021
7. McNicol E, Ferguson M, Bungay K, Rowe EL, Eldabe S, Gewandter JS, Hayek SM, Katz N, Kopell BH, Markman J, et al. Systematic Review of Research Methods and Reporting Quality of Randomized Clinical Trials of Spinal Cord Stimulation for Pain. *J Pain* (2021) **22**:127–142. doi: 10.1016/j.jpain.2020.05.001
8. Dechartres A, Trinquart L, Atal I, Moher D, Dickersin K, Boutron I, Perrodeau E, Altman DG, Ravaud P. Evolution of poor reporting and inadequate methods over time in 20 920 randomised controlled trials included in Cochrane reviews: research on research study. *BMJ* (2017)j2490. doi: 10.1136/bmj.j2490
9. Danilla S, Wasiak J, Searle S, Arriagada C, Pedreros C, Cleland H, Spinks A. Methodological quality of randomised controlled trials in burns care. A systematic review. *Burns J Int Soc Burn Inj* (2009) **35**:956–961. doi: 10.1016/j.burns.2009.04.031
10. Luchini C, Veronese N, Nottegar A, Shin JI, Gentile G, Granziol U, Soysal P, Alexinschi O, Smith L, Solmi M. Assessing the quality of studies in meta-research: Review/guidelines on the most important quality assessment tools. *Pharm Stat* (2021) **20**:185–195. doi: 10.1002/pst.2068
11. Cai X, Gewandter JS, He H, Turk DC, Dworkin RH, McDermott MP. Estimands and missing data in clinical trials of chronic pain treatments: advances in design and analysis. *Pain* (2020) **161**:2308–2320. doi: 10.1097/j.pain.0000000000001937
12. Latronico N, Botteri M, Minelli C, Zanotti C, Bertolini G, Candiani A. Quality of reporting of randomised controlled trials in the intensive care literature. A systematic analysis of papers published in Intensive Care Medicine over 26 years. *Intensive Care Med* (2002) **28**:1316–1323. doi: 10.1007/s00134-002-1339-x
13. Clark HD, Wells GA, Huët C, McAlister FA, Salmi LR, Fergusson D, Laupacis A. Assessing the quality of randomized trials: reliability of the Jadad scale. *Control Clin Trials* (1999) **20**:448–452. doi: 10.1016/s0197-2456(99)00026-4
14. Oremus M, Wolfson C, Perrault A, Demers L, Momoli F, Moride Y. Interrater reliability of the modified Jadad quality scale for systematic reviews of Alzheimer's disease drug trials. *Dement Geriatr Cogn Disord* (2001) **12**:232–236. doi: 10.1159/000051263
15. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, McQuay HJ.

Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials* (1996) **17**:1–12. doi: 10.1016/0197-2456(95)00134-4
16. Bhandari M, Richards RR, Sprague S, Schemitsch EH. Quality in the reporting of randomized trials in surgery: is the Jadad scale reliable? *Control Clin Trials* (2001) **22**:687–688. doi: 10.1016/s0197-2456(01)00147-7
17. Oremus M, Oremus C, Hall GBC, McKinnon MC, ECT & Cognition Systematic Review Team. Inter-rater and test-retest reliability of quality assessments by novice student raters using the Jadad and Newcastle-Ottawa Scales. *BMJ Open* (2012) **2**:e001368. doi: 10.1136/bmjopen-2012-001368
18. Chinea A. "Understanding the Principles of Recursive Neural Networks: A Generative Approach to Tackle Model Complexity.," In: Alippi C, Polycarpou M, Panayiotou C, Ellinas G, editors. *Artificial Neural Networks – ICANN 2009*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg (2009). p. 952–963 doi: 10.1007/978-3-642-04274-4_98
19. Lopez MM, Kalita J. Deep Learning applied to NLP. *ArXiv170303091 Cs* (2017) http://arxiv.org/abs/1703.03091 [Accessed September 27, 2021]
20. Ebrahimi J, Dou D. Chain Based RNN for Relation Classification. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics (2015). p. 1244–1249 doi: 10.3115/v1/N15-1133
21. Ren Y, Fei H, Peng Q. Detecting the Scope of Negation and Speculation in Biomedical Texts by Using Recursive Neural Network. *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Madrid, Spain: IEEE (2018). p. 739–742 doi: 10.1109/BIBM.2018.8621261
22. Good P, Haywood A, Gogna G, Martin J, Yates P, Greer R, Hardy J. Oral medicinal cannabinoids to relieve symptom burden in the palliative care of patients with advanced cancer: a double-blind, placebo controlled, randomised clinical trial of efficacy and safety of cannabidiol (CBD). *BMC Palliat Care* (2019) **18**:110. doi: 10.1186/s12904-019-0494-6
23. McGuire P, Robson P, Cubala WJ, Vasile D, Morrison PD, Barron R, Taylor A, Wright S. Cannabidiol (CBD) as an Adjunctive Therapy in Schizophrenia: A Multicenter Randomized Controlled Trial. *Am J Psychiatry* (2018) **175**:225–231. doi: 10.1176/appi.ajp.2017.17030325
24. Fitzgerald A, Mac Giollabhui N, Dolphin L, Whelan R, Dooley B. Dissociable psychosocial profiles of adolescent substance users. *PloS One* (2018) **13**:e0202498. doi: 10.1371/journal.pone.0202498
25. Anthenelli RM, Benowitz NL, West R, St Aubin L, McRae T, Lawrence D, Ascher J, Russ C, Krishen A, Evins AE. Neuropsychiatric safety and efficacy of varenicline, bupropion, and nicotine patch in smokers with and without psychiatric disorders (EAGLES): a double-blind, randomised, placebo-controlled clinical trial. *Lancet Lond Engl* (2016) **387**:2507–2520. doi: 10.1016/S0140-6736(16)30272-0
26. Avidan MS, Maybrier HR, Abdallah AB, Jacobsohn E, Vlisides PE, Pryor KO, Veselis RA, Grocott HP, Emmert DA, Rogers EM, et al. Intraoperative ketamine for prevention of postoperative delirium or pain after major surgery in older adults: an international, multicentre, double-blind, randomised clinical trial. *Lancet Lond Engl* (2017) **390**:267–275. doi: 10.1016/S0140-6736(17)31467-8
27. Yu J, Wells J, Wei Z, Fewtrell M. Effects of relaxation therapy on maternal psychological state, infant growth and gut microbiome: protocol for a randomised controlled trial investigating mother-infant signalling during lactation following late preterm and early term delivery. *Int Breastfeed J* (2019) **14**:50. doi: 10.1186/s13006-019-0246-5
28. Karaglani E, Thijs-Verhoeven I, Gros M, Chairistanidou C, Zervas G, Filoilia C, Kampani T-M, Miligkos V, Matiatou M, Valaveri S, et al. A Partially Hydrolyzed Whey Infant Formula Supports Appropriate Growth: A Randomized Controlled Non-Inferiority Trial. *Nutrients* (2020) **12**:E3056. doi: 10.3390/nu12103056
29. Bouadma L, Luyt C-E, Tubach F, Cracco C, Alvarez A, Schwebel C, Schortgen F, Lasocki S, Veber B, Dehoux M, et al. Use of procalcitonin to reduce patients' exposure

to antibiotics in intensive care units (PRORATA trial): a multicentre randomised controlled trial. *Lancet Lond Engl* (2010) **375**:463–474. doi: 10.1016/S0140-6736(09)61879-1
30. Hino H, Oda Y, Yoshida Y, Suzuki T, Shimada M, Nishikawa K. Electrophysiological effects of desflurane in children with Wolff-Parkinson-White syndrome: a randomized crossover study. *Acta Anaesthesiol Scand* (2018) **62**:159–166. doi: 10.1111/aas.13023
31. Chaitman BR, Pepine CJ, Parker JO, Skopal J, Chumakova G, Kuch J, Wang W, Skettino SL, Wolff AA, Combination Assessment of Ranolazine In Stable Angina (CARISA) Investigators. Effects of ranolazine with atenolol, amlodipine, or diltiazem on exercise tolerance and angina frequency in patients with severe chronic angina: a randomized controlled trial. *JAMA* (2004) **291**:309–316. doi: 10.1001/jama.291.3.309
32. Olivo SA, Macedo LG, Gadotti IC, Fuentes J, Stanton T, Magee DJ. Scales to Assess the Quality of Randomized Controlled Trials: A Systematic Review. *Phys Ther* (2008) **88**:156–175. doi: 10.2522/ptj.20070147
33. Foulquier N, Redou P, Le Gal C, Rouvière B, Pers J-O, Saraux A. Pathogenesis-based treatments in primary Sjogren's syndrome using artificial intelligence and advanced machine learning techniques: a systematic literature review. *Hum Vaccines Immunother* (2018)1–6. doi: 10.1080/21645515.2018.1475872
34. Foulquier N, Rouvière B, Saraux A. Can we use artificial intelligence for systematic literature review in rheumatology? *Joint Bone Spine* (2021) **88**:105109. doi: 10.1016/j.jbspin.2020.105109

Tables

Item number	question	yes	no

1	Is the clinical trial randomised ?	1	0
2	Is the clinical trial double-blinded ?	1	0
3	Does the clinical trial detail withdrawals from the study and their reasons ?	1	0
4	Does the clinical trial detail the randomisation carried out ?	1	0
5	Does the clinical trial detail the blinding performed ?	1	0

Table 1 : Description of the items used to compute a simplified Jadad score and their associated values

Item	Item 1	Item 2	Item 3	Item 4	Item 5
ACC	0,93	0,98	0,97	0,96	0,97
AUC	0,97	0,97	0,95	0,96	0,95

Table 2: Accuracy and AUC values for each of the RNN models on the first validation dataset.

Clinical trial	Items automatically predicted	Score automatically predicted	Items manually predicted	Score manually predicted
1 (22)	1 - 2	2	1 - 2	2
2 (23)	1 - 2	2	1 - 2 - 4	3
3 (24)	0	0	0	0
4 (25)	1 - 2 - 4 - 5	4	1 - 2 - 4 - 5	4
5 (26)	1 - 2 - 5	3	1 - 2 - 4 - 5	4
6 (27)	1	1	1	1
7 (28)	1 - 2 - 3	3	1 - 2 - 3	3
8 (29)	1 - 3	2	1 - 3 - 4	3
9 (30)	1	1	1	1
10 (31)	1 - 2	2	1 - 2	2

Table 3: Results obtained with the second validation dataset, grey lines highlight omitted items.

Figures

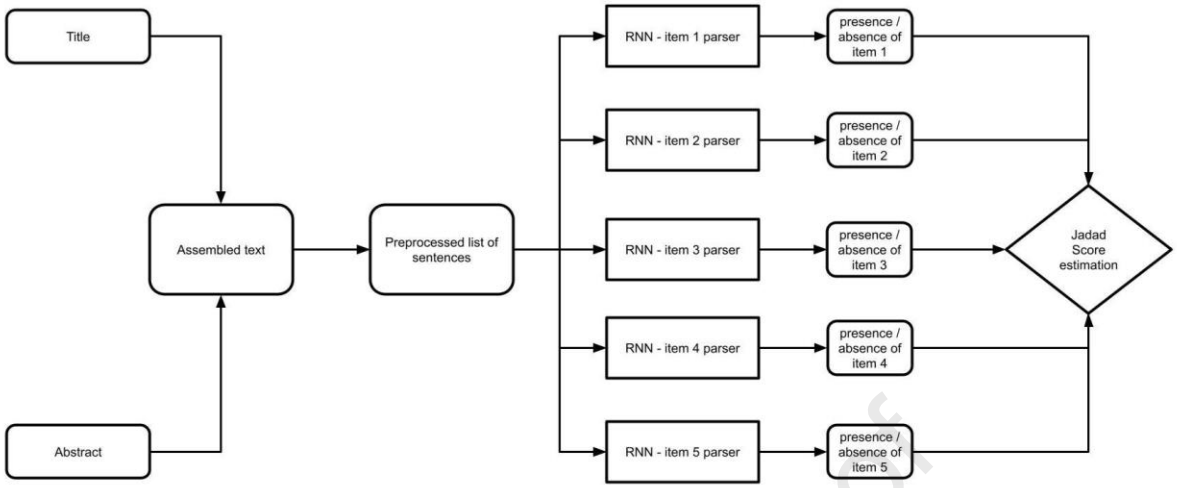


Figure 1: Representation of the program structure : the title and the abstract are first assembled into a raw text, then processed (removal of stop words, tokenization) and divided into sentences. Sentences are submitted to each of the five RNN trained to spot a specific item. Output of the RNNs are finally assembled into one final score.

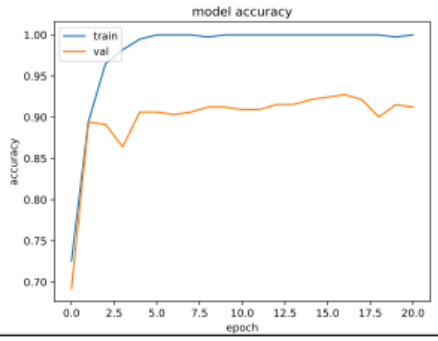
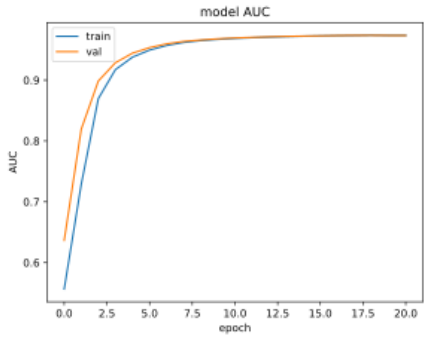
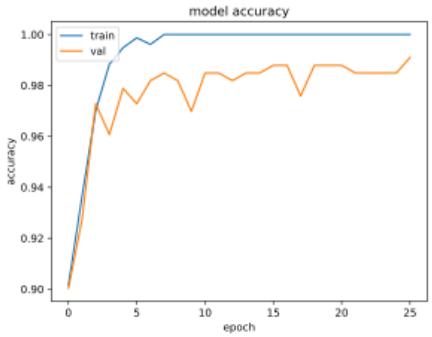
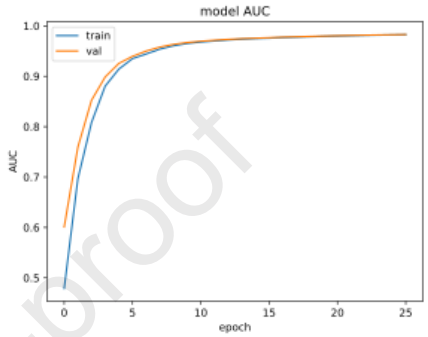
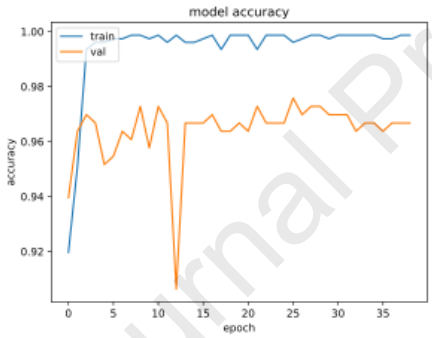
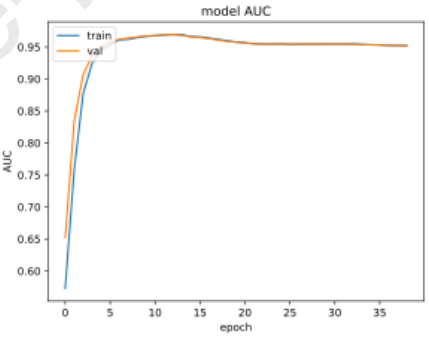
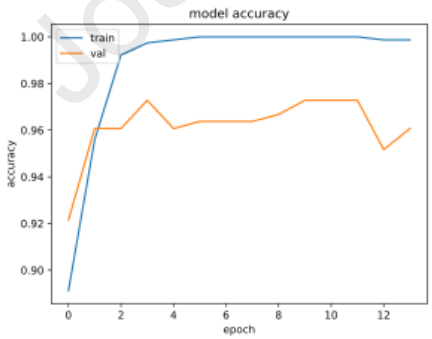
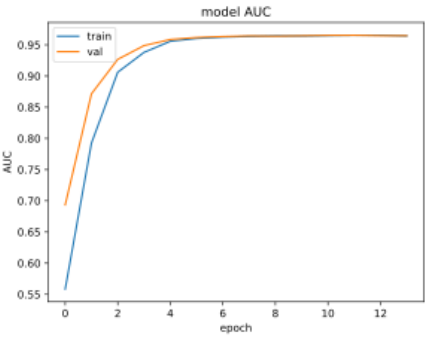
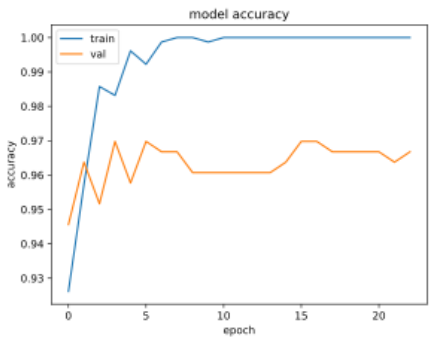
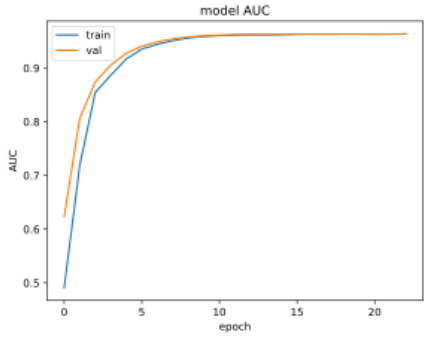
Item 1		
Item 2		
Item 3		
Item 4		
Item 5		

Figure 2: Accuracy and AUC obtained for each epoch and each of the RNN models on the first validation dataset

Assessing the robustness of clinical trials by estimating Jadad's score using artificial intelligence approaches

Tiphaine Casy (1), Alexis Grasseau (2), Amandine Charras (3) Bénédicte Rouvière (4)(5) Jacques-Olivier Pers (5), Nathan Foulquier *(5) & Alain Saraux *(5)(6)

Highlight

- We built a program based on artificial intelligence approaches to assess the robustness of a clinical trial using the Jadad score.
- The program is composed of five Recursive Neural Networks (RNN), each of them trained to spot one specific item constituting the Jadad's scale.
- After training, the algorithm had an excellent accuracy on two distinct validation sets

Prof. Alain Saraux
Service de Rhumatologie, CHU de la Cavale-Blanche,
Boulevard Tanguy Prigent,
Brest 29200
France
Tel: +33 298 22 3333
Email: alain.saraux@chu-brest.fr

Artificial Intelligence in Medicine
17 July 2022

Assessing the robustness of clinical trials by estimating Jadad's score using artificial intelligence approaches

Tiphaine Casy (1), Alexis Grasseau (2), Amandine Charras (3) Bénédicte Rouvière (4)(5) Jacques-Olivier Pers (5), Nathan Foulquier *(5) & Alain Saraux *(5)(6)

Declarations of interest: none