# Overview of the 2021 ALTA Shared Task: Automatic Grading of Evidence, 10 years later

**Diego Mollá**
Department of Computing
Macquarie University
diego.molla-aliod@mq.edu.au

## Abstract

The 2021 ALTA shared task is the 12th instance of a series of shared tasks organised by ALTA since 2010. Motivated by the advances in machine learning in the last 10 years, this year's task is a re-visit of the 2011 ALTA shared task. Set within the framework of Evidence Based Medicine (EBM), the goal is to predict the quality of the clinical evidence present in a set of documents. This year's participant results did not improve over those of participants from 2011.

## 1 Introduction

Evidence Based Medicine (EBM) urges the medical practitioner to make use of the best available evidence for making decisions about the care of individual patients (Sackett et al., 1996). However, medical and biomedical research generates such a volume of publications that it is unrealistic for a medical doctor or researcher to be able to read all relevant publications in order to be up to date on the available medical evidence. For example, PubMed currently contains more than 33 million citations for biomedical literature[1]. A more recent collection, CORD-19, contains over 500,000 publications on topics related to COVID-19, SARS-CoV-2, and related coronaviruses[2].

An important step for determining the best clinical evidence is to grade the quality of the available evidence. To help address this problem, in 2011 the ALTA shared task launched the task of automatic evidence grading (Mollá and Sarker, 2011). The goal of the task was to build a system that predicts the grade of evidence available in a set of medical publications. Forward 10 years, in 2021, the task has been re-visited. The 2021 task uses the same training and test data sets as in 2011, and the evaluation framework has been re-created as closely as possible to match the 2011 evaluation framework.

We wanted to know whether the recent advances in machine learning over the last 10 years lead to an improvement in the accuracy of the automatic grading of evidence predictors. This paper describes the specific set up of the 2021 ALTA shared task, and shows the results of the participating systems. Back in 2011, no participating systems improved on a majority baseline. In 2021, the results of the participating systems appear to improve over the majority baseline, but the difference is not statistically significant. Section 2 gives more details about the automatic grading of evidence task. Section 3 presents related work since 2011. Section 4 details the evaluation framework. Section 5 presents the participating systems and their results, and Section 6 concludes this paper.

## 2 Evidence Grading

Several taxonomies have been defined to grade the quality of the medical evidence. The Strength of Recommendation Taxonomy (SORT) (Ebell et al., 2004), used in the 2011 ALTA shared task, is one such taxonomy. SORT uses a 3-point scale defined as follows:

**A** Recommendation based on consistent and good quality patient-oriented evidence.

**B** Recommendation based on inconsistent or limited quality patient-oriented evidence.

**C** Recommendation based on consensus, usual practice, opinion, disease-oriented evidence, and case series for studies of diagnosis, treatment, prevention, or screening.

In addition to the above definitions, Ebell et al. (2004) provides details on how to determine each grade, including a flowchart.

[1] https://pubmed.ncbi.nlm.nih.gov/

[2] https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

Medical evidence is not necessarily bound to one publication only. There may be several publications related to a particular disease, treatment or diagnosis, and each of them may be of different quality. Further, it may indeed happen that each of the separate publications produces consistent results, but the evidence of the set of publications is inconsistent; when that happens the evidence grade cannot be of type A, as per the definitions above.

## 3 Related Work

The 2011 ALTA shared task overview paper (Mollá and Sarker, 2011) presents a short survey of related work prior to 2011. As we see in this section, there has been limited research since then.

None of the participants to the 2011 ALTA shared task (Mollá and Sarker, 2011) outperformed a majority baseline ("predict B", with an accuracy of 0.4863), and the participating systems did not publish the system descriptions.

A more sophisticated approach developed by the organisers of the 2011 shared task did manage to beat the baseline, reaching an accuracy of 0.6284. Their approach was based on cascaded Support Vector Machine (SVM) classifiers which were trained to separate class A and C from the default B with high precision. These SVM classifiers used combinations of the following features: $n$-grams of the abstract and title (with general medical semantic types replacing specific medical terms), and publication types (combining the publication types provided in the original abstracts with types generated by applying *ad-hoc* rules). The work was subsequently extended and published with more detail by Sarker et al. (2015).

Gyawali et al. (2012) reported an improved accuracy of 0.7377 on the same dataset by using a two-level stacking approach. In the first level, multiple SVM classifiers are trained using separate feature sets. Then, their output is fed to a second SVM classifier. Their feature sets included publication types, MeSH terms, title, abstract text, abstract method section, and abstract conclusion section. All of these features were as provided by the abstracts, except for the method and conclusion section, which were determined heuristically when not provided by the abstracts.

Byczyńska et al. (2020) reported an accuracy of 0.7541, again on the same dataset, after applying a wide range of different variants of stacked classifiers.

```
00001 B 10553790 15265350
00002 C 12804123 16026213 14627885
00003 B 15213586
00004 A 15329425 9058342 11279767
```

Figure 1: Sample training data. Each row indicates one evidence that needs to be graded. The first number is the evidence ID. This is followed by the evidence grade, and the list of PubMed IDs for the relevant documents.

Table 1 shows the results of the works mentioned in this section, with their confidence intervals as calculated by the Wilson score interval with continuity correction (Brown et al., 2001). According to the confidence intervals shown on the table, the difference between the systems by Gyawali et al. (2012) and Byczyńska et al. (2020) is not statistically significant.

## 4 Evaluation Framework

The data for the 2021 shared task includes a training set and a development set that were available to the participants. The final ranking was made on a separate test set and was available to the participants (without the target labels) for a limited time near the end of the shared task.

The training, development, and test sets were the same as for the 2011 shared task, after shuffling the rows and changing the row IDs. The corpus from which this data has been obtained has been described by Mollá et al. (2016). Figure 1 illustrates a fragment of the training data. Together with the data formatted as the samples of Figure 1, the participants were provided with the contents of the relevant abstracts as separate files.

The evaluation framework was implemented as a CodaLab competition[3]. The facilities available at CodaLab made it possible to specify our own evaluation script, and also gave us flexibility to design multiple phases and include a leaderboard and discussion forum. Additional information about the 2021 ALTA shared task was made available in the ALTA website[4].

The CodaLab competition was structured into two phases. In a first, development phase, all teams had access to the training and development sets and they could make an arbitrary number of submissions daily, for a maximum of 100 submissions in total. During the development phase, participant teams could submit the results of running their

[3] https://competitions.codalab.org/competitions/33739

[4] http://www.alta.asn.au/events/sharedtask2021/

| System | Accuracy | 95% CI |
|---|---|---|
| Majority Baseline | 0.4863 | 0.4150–0.5583 |
| Mollá and Sarker (2011) | 0.6284 | 0.5564–0.6951 |
| Gyawali et al. (2012) | 0.7377 | 0.6696–0.7961 |
| Byczyńska et al. (2020) | 0.7541 | 0.6869–0.8108 |

Table 1: Accuracy and 95% confidence intervals of prior work. The confidence intervals were calculated using the Wilson score interval with continuity correction.

system on the development data, and the results could enter a public leaderboard. In the second, test phase, all teams had access to the test data set and each team could make a maximum of 3 submissions. The final ranking was made based on the best submission of each team made during the test phase. Table 2 shows the timeline and submission number limits of each phase.

The evaluation metric was accuracy.

## 5 Participating Systems

As in past ALTA shared tasks, submissions were made by teams in two categories: a student category, and an open category. In teams of the student category, all members must be university students and none of the team members could have a PhD. Teams that did not qualify for the student category could participate in the open category.

A total of 16 teams registered in the student category, and 5 teams registered in the open category. Of these, only 5 teams, all from the student category, submitted runs in the test phase for final ranking.

Table 3 shows the results of the systems by the participating teams. As can be observed, none of them improves the upper confidence interval of the majority baseline (0.5583). A McNemar's test for statistical significance confirmed that none of the submitted systems had a statistically significant difference with the majority baseline.

Of the 5 teams submitting in the final phase, 3 published a system description which is available in the 2021 ALTA proceedings. Team SarkerLab (Guo et al., 2021) experimented with the use of SVM and RoBERTa. Team Heatwave (Koto and Fang, 2021) applied an ensemble method with transformer variants including BioMed, RoBERTa, and ELECTRA. Finally, team OrangUtanV3 (Parameswaran et al., 2021) applied a cascaded approach that used BioBERT and SVM classifiers. Whereas team Heatwave's classifiers attempted to generate the final evidence grade of the collection of abstracts related to a question, the other two teams attempted to classify individual abstracts and the final result was obtained by combining the outputs of the individual classifications.

## 6 Conclusions

The participating systems appeared to obtain a score slightly better than the majority baseline but the difference was not statistically significant. These results underperformed those reported by the organisers of the 2011 shared task paper and subsequent work. The participating systems attempted to use some of the latest developments on machine learning algorithms and architectures. The reason of their relatively lower performance may be due to the choice of features. Possibly, better results could have been obtained by incorporating information such as the publication type, or by focusing on specific parts of the abstracts such as the methods or conclusions sections, as related work has shown to be most influential for this task.

## References

Lawrence D. Brown, T. Tony Cai, and Anirban DasGupta. 2001. Interval Estimation for a Binomial Proportion. *Statistical Science*, 16(2):101 – 133.

Aleksandra Byczyńska, Maria Ganzha, Marcin Paprzycki, and Mikołaj Kutka. 2020. Evidence quality estimation using selected machine learning approaches. In *2020 Conference on Information Communications Technology and Society (ICTAS)*, pages 1–8.

Mark H Ebell, Jay Siwek, Barry D Weiss, Steven H Woolf, Jeffrey Susman, Bernard Ewigman, and Marjorie Bowman. 2004. Strength of recommendation taxonomy (SORT): a patient-centered approach to grading evidence in the medical literature. *The Journal of the American Board of Family Practice / American Board of Family Practice*, 17:59–67.

Yuting Guo, Yao Ge, Ruqi Liao, and Abeed Sarker. 2021. An ensemble model for automatic grading of evidence. In *Proceedings of the 2021 Australasian Language Technology Workshop*.

| Phase | From | To | Submissions |
|---|---|---|---|
| Development | June 30, 2021 | October 3, 2021 | 100 |
| Test | October 4, 2021 | October 11, 2021 | 3 |

Table 2: Dates and submission number limits of the CodaLab competition phases.

| Rank | Team | Accuracy |
|---|---|---|
| 1 | SarkerLab | 0.5355 |
| 2 | Heatwave | 0.5027 |
| 3 | OrangUtanV3 | 0.4918 |
| 4 | arana-initiatives | 0.4863 |
|  | (Majority Baseline | 0.4863) |
| 5 | nikss | 0.4536 |

Table 3: Results of the participating systems plus the majority baseline.

Binod Gyawali, Thamar Solorio, and Yassine Benajiba. 2012. Grading the quality of medical evidence. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 176–184, Montréal, Canada. Association for Computational Linguistics.

Fajri Koto and Biaoyan Fang. 2021. Handling variance of pretrained language models in grading evidence in the medical literature. In *Proceedings of the 2021 Australasian Language Technology Workshop*.

Diego Mollá, María Elena Santiago-Martínez, Abeed Sarker, and Cécile Paris. 2016. A corpus for research in text processing for evidence based medicine. *Language Resources and Evaluation*, 50:705–727.

Diego Mollá and Abeed Sarker. 2011. Automatic grading of evidence: the 2011 ALTA shared task. pages 4–8. Australian Language Technology Association.

Pradeesh Parameswaran, Andrew Trotman, Veronica Liesaputra, and David Eyers. 2021. Quick, get me a Dr. BERT: Automatic grading of evidence using transfer learning. In *Proceedings of the 2021 Australasian Language Technology Workshop*.

David L. Sackett, William M. Rosenberg, Jamuir Gray, R. Brian Haynes, and W. Scott Richardson. 1996. Evidence based medicine: What it is and what it isn't. *BMJ*, 312:71–72.

Abeed Sarker, Diego Mollá, and Cécile Paris. 2015. Automatic evidence quality prediction to support evidence-based decision making. *Artificial Intelligence in Medicine*, 64:89–103.