Quick, get me a Dr. BERT: Automatic Grading of Evidence using Transfer Learning

Pradeesh Parameswaran, Andrew Trotman, Veronica Liesaputra, David Eyers
Department of Computer Science
University of Otago
New Zealand
[pradeesh,andrew,veronica,dme]@cs.otago.ac.nz

Abstract

We describe our methods for automatically grading the level of clinical evidence in medical papers, as part of the ALTA 2021 shared task. We use a combination of transfer learning and a hand-crafted, feature-based classifier. Our system ("*orangutanV3*") obtained an accuracy score of 0.4918, which placed third in the leaderboard. From our failure analysis, we find that our classification techniques do not appropriately handle cases when the conclusions of across the medical papers are themselves inconclusive. We believe that this shortcoming can be overcome—thus improving the classification accuracy—by incorporating document similarity techniques.

1 Introduction

The recent COVID-19 pandemic has once again highlighted the importance of Evidence-Based Medicine (EBM) when deciding the course of treatment (Xu et al., 2020) as social media and television shows are being flooded by so-called experts, who have voiced unproven treatments for COVID-19 such as using hydroxychloroquine (Greenhalgh, 2020; Aquino and Cabrera, 2020). However, the main challenge with EBM is that it is a manual and tedious process and it is very hard for practitioners to keep up with the rise in medical research (Ghosh, 2004; Davies, 2007).

The challenges with EBM were no less true in 2011, when the Australasian Language Technology Association (ALTA) organised a shared task challenge to automatically grade evidence (Molla and Sarker, 2011). The task was to grade evidence based on an EBM framework which is the Strength Of Recommendation Taxonomy (SORT) (Ebell et al., 2004). ALTA decided to revisit the 2011 challenge again this year (2021), motivated by the leaps in Natural Language Processing (NLP) techniques that have occurred meanwhile (Torfi et al., 2020).

We investigate the following research questions with repsect to this challenge:

- *RQ1*: Can we solely use transformers to accurately perform SORT?
- *RQ2*: Can we improve the performance of Transformers by incorporating author and journal features?

Our experimental results suggest that these two approaches do not perform well. Our team placed third in the leaderboard with an accuracy score of 0.4918. None of the scores on the leaderboard beat the winning accuracy score in 2011, of 0.6284. This prompted us to perform an in-depth analysis of our approach, and how our work can be improved in the future to increase the overall accuracy of classification.

2 Related Work

In the medical literature, there are many different taxonomies that are used in order to rank the grade of a clinical study (Abrams et al., 2007; Guyatt et al., 2004). One of the commonly used taxonomies, due to its simplicity, is SORT (Ebell et al., 2004). SORT has been used in deciding whether to recommend root canal treatments (De-Deus and Canabarro, 2017), sports injury rehabilitation strategies (Bell et al., 2018; Rodriguez et al., 2019), and in evaluation of cognitive behavioural treatment (Chang et al., 2020; Baez et al., 2018). There are three grades: `A` (strong), `B` (moderate) and `C` (weak). Grade `A` reflects consistent and good-quality, patient-oriented evidence; Grade `B` reflects being based on inconsistent or limited quality patient oriented evidence; lastly, grade `C` reflects a recommendation based on consensus, usual practice, opinion or disease-oriented evidence.

The classification of SORT is manually done by medical practitioners, and automating it is still in its infancy. To the best of our knowledge, the only researchers who explored automating SORT are Molla and Sarker (2011). In their work, the authors used a set of classifiers that utilised different feature sets such as n-grams, publication type and titles and then applied multiple SVM classifiers. They obtained an accuracy score of 0.6284.

Transfer learning has shown vast improvement on a variety of downstream tasks such as summarization, translation, and question and answer interactions (Torfi et al., 2020). One popular transfer learning method that is widely adapted is BERT (Devlin et al., 2018). Driven by the success of BERT, Lee et al. (2020) introduced BioBERT (a biomedical focused version of BERT) for tasks such as biomedical Named Entity Recognition (NER), relation extraction, and summarization in the biomedical literature. Recently, Oniani and Wang (2020) demonstrated that BioBERT provides an effective method for chatbots answering questions related to COVID-19.

3 Data Set

The data set[1] provided by the organisers of the ALTA shared task consists of a collection of PubMed abstracts. There are 677 medical abstracts for training, 178 for development, and 183 for the testing set. The training and development data set come with the evidence ID, followed by SORT grade, and finally a list of PubMed IDs of the abstracts. The test data contains the same, except for the SORT grade.

We analysed both the development and training data sets to understand the characteristics of the data. Table 1 shows the distribution of the evidence that contains exactly one abstract and more than one abstract. We include the class distribution of both training and development sets. Across three data sets on average, the percentage of evidence IDs that contains more than one PubMed ID are 57%. From the visual inspection of our training set, we have observed that the majority of the queries (77%) tend to be graded as `A` and `B`. We have also noticed that the distribution of classes in development follows closely that of the training set.

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

Data Set	= 1 abstract	> 1 abstract	No of A	No of B	No of C
Train	293	384	212	311	154
Dev	113	65	48	80	50
Test	105	78	NA	NA	NA

Table 1: Distribution and abstracts in the data sets.

4 Methodology

We employed a two-phase approach to tackle the ALTA challenge. In the first phase, we used a pre-trained BioBERT model and in the second phase, we used an SVM classifier with handcrafted features such as h-index and the journal's impact factor. In this section, we describe our method in detail, along with the steps that we performed. We have made our system's source code publicly available on GitHub.[2]

4.1 Phase 1—BioBERT

We used a pre-trained BioBERT model *biobert-base-cased-v1.2*[3]. The two primary reasons for choosing this model is that the implementation is readily available via *huggingface*,[4] and that it has been trained on PubMed. Since our task relates to grading medical abstracts which are obtained from PubMed, this gives us further confidence that BioBERT would be the right choice for our task.

We first extract the abstracts using the PubMed IDs. If there are multiple PubMed IDs for a piece of given evidence, we treat each of them as independent from one another. This made the implementation easier. We then pre-processed the texts with *Scispacy*[5] by replacing entities of diseases with `[DISEASE]`, drug names as `[DRUGS]` and treatment plans with `[TREATMENT]`. Replacing these instances with a generic tag ensures that the classifier does not overfit or get influenced by these factors. We used the same pre-trained model for pre-processing. In addition to that, we replaced instances of sample size conducted in the studies by following the recommendation from Biau et al. (2008); Charan and Biswas (2013) into three generic tags: `[SMALL]` when the sample size is less than 15, `[MEDIUM]` when it is between 15–100, and `[LARGE]` when it is greater than 100.

[2] https://github.com/prasys/OrangUtanV3ALTASharedTask21

[3] https://huggingface.co/dmis-lab/biobert-base-cased-v1.2

[4] https://huggingface.co/

[5] https://allenai.github.io/scispacy/

These additional steps were done in order to prevent the model from over-fitting.

We then built two classifiers of BioBERT. The first classifier (C_1) distinguishes C-graded documents from non-C graded documents. As for the non-C graded documents, it is then fed to our second BioBERT classifier (C_2) which distinguishes A-graded documents from B-graded documents.

We evaluated the performance of our classifier using the validation data set. As for fine-tuning the classifier, we split 80% of our training data for training and the remaining 20% for fine-tuning. We froze all the layers of the model except for the final layer which is the classification layer. We used the Adam Optimizer with a learning rate of 10^{-5} for 10 epochs for C_1 and a learning rate of $10^{-3}$for 15 epochs for C_2. We set the batch size to be 64 for both C_1 and C_2. For both the classifiers, we used our validation set's accuracy score as an early stopping criterion. We stop the training if the score does not increase for 5 consecutive epochs or the maximum number of epochs has been reached. Our model was entirely implemented using huggingface 4.6.1.

4.2 Phase 2—SVM

In the second phase, we use an SVM classifier (C_3) with the following feature set; authors' h-index (averaged across all of the authors), journal's impact factor and also the journal rank. Past studies (Lee et al., 2002; Saha et al., 2003) have shown that these criteria can be used to judge the quality of medical literature and we hypothesise that these would further help to distinguish B graded articles from A graded articles given that the criteria for these grades are finer.

We made several assumptions when we derived our features. Firstly, given that the data set contains articles published from the late 1980s up to the late 2000s, the journal name at the time of publication may have changed. To tackle this problem, we obtained the current journal name using *google-scholar-crawler*[6] to crawl Google Scholar in order to retrieve the journal's current name, along with each author's h-index. This took us a considerable amount of time, as we were being rate-limited by the number of queries that Google allows, and to the best of our knowledge, there aren't any publicly available APIs for us to use.

[6] https://github.com/geekan/google-scholar-crawler

System	Accuracy Score
$C_1 + C_2$	0.4494
$C_1 + C_3$	0.4228
$C_1 + C_2 + C_3$	**0.6573**

Table 2: System evaluation on development set.

System	Accuracy Score
$C_1 + C_2$	0.4808
$C_1 + C_3$	**0.5010**
$C_1 + C_2 + C_3$	0.4918

Table 3: System evaluation on test set.

We took the journal ranking and the impact factor from the 2020 SCImago[7] because we were unable to obtain the journals ranking and impact factor at the time of publication of the article.

We use sklearn 0.9.4[8] for implementation and kept the default kernel parameters of the SVM classifier as it yielded the best results based on our early experimentation. We used the same split as that in our phase one classifier.

4.3 Final Prediction

To make the final prediction, first we ran the abstracts through C_1 and then for those classified as non-C, we ran through C_2 and finally we ran the same set again through C_3. We set C_3 to have higher precedence than C_2. If a piece of evidence is categorised as A by C_2 and as B as C_1, we set the grade to B. If there are multiple abstracts for a given evidence ID, we assign a score of 3 for A, 2 for B and 1 for C. We then calculate the mean score and assign the grade closest to the score.

5 Results

ALTA chose CodaLabs as the submission platform. The organisers provided us with both the development and test set. In CodaLabs, participants are allowed to submit an unlimited number of times for the development set but are only limited to three submissions for the test set. The submissions are evaluated by using the accuracy score. We ran our experiments on Google Cloud Platform with 4 vCPUs, 16 GiB of RAM and an NVIDIA Tesla A100. We present our scores on the development set in Table 2 and summarise and present our final results on the test set in Table 3.

[7] https://www.scimagojr.com/

[8] https://scikit-learn.org/stable/

5.1 Discussion

From our experimental results on the development set, we found that our approach of using $C_1 + C_2 + C_3$ yielded the best results with a score of 0.6573 and a two-way ANOVA test confirmed that there are statistically significant differences between the systems ($p < 0.05$). Therefore, we picked $C_1 + C_2 + C_3$ to be used in the test portion.

When we evaluated our system on the test portion, we were surprised to see that we obtained a score of 0.4918. This suggested that our models are most likely over-trained or have not learnt to generalise very well. We went back to the drawing board to see if we could further improve the scores. Given the limited number of submissions, we decided to submit the other two models to see if they would fare better. To our surprise, $C_1 + C_3$ gave us the best results with a score of 0.5010. A one-way ANOVA showed no statistically significant difference between the scores at the 0.05 level and so we decided to look deeper at $C_1 + C_2 + C_3$ to have an understanding of what went wrong. We discuss this in subsection 5.2, 5.3 and 5.4.

5.2 Ambiguity in Classifying Different Grades

We first evaluated the performance of our three proposed systems ($C_1 + C_2$, $C_1 + C_3$, and $C_1 + C_2 + C_3$) on the development set containing only a single abstract and found that the best performing classifier is $C_1 + C_2$ with an accuracy score of 0.8314, followed by $C_1 + C_2 + C_3$ with an accuracy score of 0.8167, and finally $C_1 + C_3$ gave us a score of 0.7854. Our one-way ANOVA test also showed that there is a statistically significant difference ($p < 0.05$) between $C_1 + C_2 + C_3$ and the others.

Looking at the causes of failures with evidence containing a single abstract, we notice that distinguishing between A and B can be challenging. We observed that out of 13 A grades, we incorrectly classified 5 as B. As for B grades, out of 33, we misclassified 6 as A. Additionally, we have notice that 5% of the instances where C grade evidence are misclassified as A and B. We provide some examples of our findings from the development set in Table 5.

In the first example, the evidence is about diagnosing carpal tunnel syndrome. If we take a closer look at the abstract, we observe keywords that are commonly found in A and B graded evidence such as *"p-value"* , *"CI"* . However, unlike in many A and B graded papers, the authors of this paper mention that a *"future randomised control trial is required to validate the results"*. We observed that our model did not understand the context that the randomised control trial has not been performed and therefore classified it as B.

ID	Abstracts	Predicted	Actual
10111	Walker et al. (1993); Zajecka (2001); Ferguson et al. (2001)	A	B
10132	Frisancho (2000); Parsons et al. (1999); Hediger et al. (1999)	B	B

Table 4: Comparison of predictions made by our system and the actual label for evidence IDs with multiple abstracts.

This brings us to our second example, the evidence is on managing chronic fatigue syndrome in a primary care setting. Our classifier classifies it as C. From our visual inspection, we agree with the decision of the classifier. However, the annotators have graded it as B. Since none of the authors is an EBM practitioner, we cannot accurately determine the reason for this. We believe that an EBM practitioner would be able to provide us with how the decisions are derived, which can help to further improve our model. We'll leave this to be part of our future work.

As for the last example, this is an example of how authors' h-index and journal ranking influences the final grading of the evidence. Initially, the $C_1 + C_2$ classifiers labelled it as grade B but C_3 classifier classified it as A. Upon inspecting further, we find that C_3 classified it as A as the authors' h-index and the journal ranking fall in the A listing. We also observed that there are times where this information helped to correct the classification of phase one classifiers such as in evidence ID *10079* (Jackson et al., 1999) and *10042* (Orton and Omari, 2008). From our analysis, we hypothesise that these factors influence the grading of the paper in a similar way to the way funding source influences the quality of the study (Reed et al., 2007). Although SORT's assessment criteria do not mention this, our investigation suggests that this needs to be explored further.

5.3 Ablation study

To further understand the effect of the features that are used in C_3, we perform an ablation study on

the development set that contains a single abstract. Keeping it with one single abstract allows us to separate our assumption with multiple documents per query. Additionally, we decided to only tune the C_3 classifier whilst keeping the other two classifiers as they are as we are interested in the impact of how different features influences the score. We summarise our results in Table 6.

For our ablation study, we looked at several aspects. First, we looked at using the impact on the accuracy score by solely using the primary author's h-index and averaging all the authors' h-index scores. We have noticed that if we were to use the primary author's h-index instead of calculating the mean h-index score of all the authors, the score decreased from 0.6603 to 0.6327. This is mainly because in the medical literature field, generally, the last author is the grand-holder or a prominent researcher in the field (Pina et al., 2019). From our test, we find no statistically significant difference ($p < 0.05$). However, given that the number of cases containing 1 abstract is small (n = 113), we think that the statistical power is limited, thus we decided to proceed on with our decision of averaging the h-index.

Additionally, we investigated independently the impact on journal ranking and h-index. We found that these two features have a high correlation coefficient (r) score of 0.92. However, if we removed one of the features, we notice that the scores decreases from 0.8167 to 0.7015 or to 0.6716. Our experimental results suggest that even highly correlated variables could carry non-redundant information, thus removing either degrades the overall information content.

5.4 Challenges with Averaging Method

Next, we repeated the experiment again—but this time solely focusing on evidence containing more than one abstract in order to test the effectiveness of our averaging method. From our experiments, we find that the best performing classifier (C_1 + C_2 + C_3) could only obtain an accuracy score of 0.4183—which is almost half the performance of the classifier on the queries containing a single abstract. This suggests that averaging the grading of each abstract is inadequate.

The score that we obtained only provides an indicator that our assumption needs to be redefined but it does not provide insights into why our performance is higher in the development set than in the test set. To answer this question, we looked at cases where prediction matches with annotators as well as the cases in which it does not match. We provide some examples of our findings from the development set in Table 4. Given that we have a limited amount of space—we provide a citation to the paper for the readers to examine instead of the complete abstract.

Evidence ID	Abstract	Predicted	Actual
10141	Plaisance et al. (2000)	B	C
10169	Kroenke et al. (1988)	C	B
10091	D'Arcy and McGee (2000)	B	A

Table 5: Comparison of predictions made by our system and the actual label for an evidence ID with a single abstract.

For the first example, the papers describe treatments of antidepressant-related sexual dysfunction. If we follow our method, all of the papers are graded as A since they fit the criteria to be graded as such. However, it was a surprise to us as to why the annotators classified it as B. Upon closely examining the three papers, we find that these papers suggest completely different mechanisms on how to address sexual dysfunction thus bumping down the grade to be B instead of an A. This finding prompted us to look closer into the way of how the final scores are calculated.

In the second example, the three papers describe the impact of obesity in children. If these are treated as a standalone, they are ranked A, B and C individually, based on SORT. In our method, we then average the grades to produce the final grade thus giving the evidence an overall score of B—matching the annotator's grade. However, we believe that this is purely by chance as when we visually inspect the abstracts—we find that the conclusions of the studies do not agree with one another, thus placing it in the B category. A better approach could be to use a Siamese Manhattan LSTM (Mueller and Thyagarajan, 2016) or even using Word Mover (Kusner et al., 2015) document similarity measures. Incorporating reinforcement learning might be able help our model to distinguish better as well. We will explore this as part of our future work.

h-index	Journal Rank	Impact Factor	Accuracy Score
Average	✗	✗	0.6603
1st Author	✗	✗	0.6327
✗	✓	✗	0.5042
✗	✗	✓	0.5565
Average	✓	✗	0.6716
1st Author	✓	✗	0.6654
Average	✗	✓	0.7015
1st Author	✗	✓	0.6968
Average	✓	✓	**0.8167**
1st Author	✓	✓	0.7669

Table 6: Ablation study of the features features used in C_3 which includes h-index (primary author's and average across all authors), the journal rank and the impact factor.

6 Conclusion

We presented an approach to automatically grade evidence using a combination of transfer learning and a feature-based classifier. We competed in the ALTA 2021 Competition under the team name "*orangutanV3*". Despite achieving an accuracy score of 0.4918, we did not manage to beat the current state-of-the-art from ten years ago. The primary reason for our low score is attributed to our assumption of averaging the grades to obtain the final grade. As for our *RQ1*, we find that solely using a transformer on single abstracts is sufficient, as we obtained a score of 0.8314 in our development set. As for our *RQ2*, we obtained a score of 0.8167, although this gives us a lower score compared to using transformers alone. We still think that combining transformer along with the SVM classifier is a a better option. However, we do not have a high statistical power to support the claim that using two models improve the overall accuracy, as we only have a limited sample size. We plan to explore further with a larger data set as part of future work. Additionally, we plan to re-implement the technique used by (Molla and Sarker, 2011) in order to properly evaluate how our system compares, when focusing on queries with a single document.

Acknowledgements

Many thanks to Vaughan Kitchen and Lo Wei Hong for their time in proofreading this paper. We would like to thank Google Cloud, Dr Diego Molla-Aliod and the organisers for ALTA for their support.

References

P Abrams, S Khoury, and A Grant. 2007. Evidence-based medicine overview of the main steps for developing and grading guideline recommendations. *Prog Urol*, 17(3):681–4.

Yves SJ Aquino and Nicolo Cabrera. 2020. Hydroxychloroquine and covid-19: critiquing the impact of disease public profile on policy and clinical decision-making. *Journal of Medical Ethics*, 46(9):574–578.

Shelby Baez, Matthew C Hoch, and Johanna M Hoch. 2018. Evaluation of cognitive behavioral interventions and psychoeducation implemented by rehabilitation specialists to treat fear-avoidance beliefs in patients with low back pain: a systematic review. *Archives of physical medicine and rehabilitation*, 99(11):2287–2298.

David R Bell, Eric G Post, Kevin Biese, Curtis Bay, and Tamara Valovich McLeod. 2018. Sport specialization and risk of overuse injuries: a systematic review with meta-analysis. *Pediatrics*, 142(3).

David Jean Biau, Solen Kernéis, and Raphaël Porcher. 2008. Statistics in brief: the importance of sample size in the planning and interpretation of medical research. *Clinical orthopaedics and related research*, 466(9):2282–2288.

Cindy Chang, Margot Putukian, Giselle Aerni, Alex Diamond, Gene Hong, Yvette Ingram, Claudia L Reardon, and Andrew Wolanin. 2020. Mental health issues and psychological factors in athletes: detection, management, effect on performance and prevention: American medical society for sports medicine position statement—executive summary. *British journal of sports medicine*, 54(4):216–220.

Jaykaran Charan and Tamoghna Biswas. 2013. How to calculate sample size for different study designs in medical research? *Indian journal of psychological medicine*, 35(2):121.

Christopher A D'Arcy and Steven McGee. 2000. Does this patient have carpal tunnel syndrome? *Jama*, 283(23):3110–3117.

Karen Davies. 2007. The information-seeking behaviour of doctors: a review of the evidence. *Health Information & Libraries Journal*, 24(2):78–94.

G De-Deus and A Canabarro. 2017. Strength of recommendation for single-visit root canal treatment: grading the body of the evidence using a patient-centred approach. *International endodontic journal*, 50(3):251–259.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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*, 17(1):59–67.

James M Ferguson, Ram K Shrivastava, Stephen M Stahl, James T Hartford, Frances Borian, John Ieni, and Darlene Jody. 2001. Reemergence of sexual dysfunction in patients with major depressive disorder: double-blind comparison of nefazodone and sertraline. *The Journal of clinical psychiatry*, 62(1):0–0.

A Roberto Frisancho. 2000. Prenatal compared with parental origins of adolescent fatness. *The American journal of clinical nutrition*, 72(5):1186–1190.

Amit K Ghosh. 2004. On the challenges of using evidence-based information: the role of clinical uncertainty. *Journal of Laboratory and Clinical Medicine*, 144(2):60–64.

Trisha Greenhalgh. 2020. Will covid-19 be evidence-based medicine's nemesis?

Gordon Guyatt, Deborah Cook, and Brian Haynes. 2004. Evidence based medicine has come a long way.

Mary L Hediger, Mary D Overpeck, Andrea McGlynn, Robert J Kuczmarski, Kurt R Maurer, and William W Davis. 1999. Growth and fatness at three to six years of age of children born small-or large-for-gestational age. *Pediatrics*, 104(3):e33–e33.

Lisa A Jackson, Patti Benson, Vishnu-Priya Sneller, Jay C Butler, Robert S Thompson, Robert T Chen, Linda S Lewis, George Carlone, Frank DeStefano, Patricia Holder, et al. 1999. Safety of revaccination with pneumococcal polysaccharide vaccine. *Jama*, 281(3):243–248.

Kurt Kroenke, David R Wood, A David Mangelsdorff, Nancy J Meier, and John B Powell. 1988. Chronic fatigue in primary care: prevalence, patient characteristics, and outcome. *Jama*, 260(7):929–934.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Kirby P Lee, Marieka Schotland, Peter Bacchetti, and Lisa A Bero. 2002. Association of journal quality indicators with methodological quality of clinical research articles. *Jama*, 287(21):2805–2808.

Diego Molla and Abeed Sarker. 2011. Automatic grading of evidence: the 2011 ALTA shared task. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 4–8, Canberra, Australia.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

David Oniani and Yanshan Wang. 2020. A qualitative evaluation of language models on automatic question-answering for covid-19. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–9.

Lois C Orton and Aika AA Omari. 2008. Drugs for treating uncomplicated malaria in pregnant women. *Cochrane Database of Systematic Reviews*, (4).

Tessa J Parsons, Chris Power, Stuart Logan, and CD Summerbelt. 1999. Childhood predictors of adult obesity: a systematic review. *International journal of obesity*, 23.

David G Pina, Lana Barać, Ivan Buljan, Francisco Grimaldo, and Ana Marušić. 2019. Effects of seniority, gender and geography on the bibliometric output and collaboration networks of european research council (erc) grant recipients. *PLoS One*, 14(2):e0212286.

Karen I Plaisance, Suneel Kudaravalli, Steven S Wasserman, Myron M Levine, and Philip A Mackowiak. 2000. Effect of antipyretic therapy on the duration of illness in experimental influenza a, shigella sonnei, and rickettsia rickettsii infections. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 20(12):1417–1422.

Darcy A Reed, David A Cook, Thomas J Beckman, Rachel B Levine, David E Kern, and Scott M Wright. 2007. Association between funding and quality of published medical education research. *Jama*, 298(9):1002–1009.

Rosa M Rodriguez, Ashley Marroquin, and Nicole Cosby. 2019. Reducing fear of reinjury and pain perception in athletes with first-time anterior cruciate ligament reconstructions by implementing imagery training. *Journal of sport rehabilitation*, 28(4):385–389.

Somnath Saha, Sanjay Saint, and Dimitri A Christakis. 2003. Impact factor: a valid measure of journal quality? *Journal of the Medical Library Association*, 91(1):42.

Amirsina Torfi, Rouzbeh A Shirvani, Yaser Keneshloo, Nader Tavaf, and Edward A Fox. 2020. Natural language processing advancements by deep learning: A survey. *arXiv preprint arXiv:2003.01200*.

Parks W Walker, Jonathan O Cole, Elmer A Gardner, Arlene R Hughes, J Andrew Johnston, Sharyn R Batey, and Charles G Lineberry. 1993. Improvement in fluoxetine-associated sexual dysfunction in patients switched to bupropion. *The Journal of clinical psychiatry*.

Guogang Xu, Yongshi Yang, Yingzhen Du, Fujun Peng, Peng Hu, Runsheng Wang, Ming Yin, Tianzhi Li, Lei Tu, Jinlyu Sun, et al. 2020. Clinical pathway for early diagnosis of covid-19: updates from experience to evidence-based practice. *Clinical reviews in allergy & immunology*, 59(1):89–100.

John Zajecka. 2001. Strategies for the treatment of antidepressant-related sexual dysfunction. *Journal of Clinical Psychiatry*, 62:35–43.