

LLM Performance Degradation Across Pediatric Age Groups: Demographic Stratified Evaluation of Clinical NLP on Synthetic Patient Cohorts

Vinod Rufus Motani ¹

¹AI/ML Engineer, Independent Researcher, USA

Email ID : motanivinodrufus18@gmail.com

ORCID: 0009-0005-3459-0859.

[Cite this paper as:](#) Vinod Rufus Motani (2025) LLM Performance Degradation Across Pediatric Age Groups: Demographic Stratified Evaluation of Clinical NLP on Synthetic Patient Cohorts. Journal of Neonatal Surgery, 14, (32s) 11250-11256

ABSTRACT

Large language models (LLMs) have become widespread in the context of clinical natural language processing (NLP). However, their effectiveness in pediatric age groups requires further research. The present study seeks to analyze the decrease in LLM performance on the basis of the demographic-stratified analysis of clinical NLP tasks performed on synthetic pediatric patient cohorts. The research is based on a qualitative secondary approach with the usage of evidences from recent articles about pediatric artificial intelligence, clinical NLP, fairness of models, and synthetic health data. The cohorts in question have been divided into four age groups: neonatal, infant, child, and adolescent. It has been found out that there is a considerable difference in the performance of LLMs among different developmental periods. Adolescent records showed the best results in terms of accuracy as their documentation format resembles the documentation format of adult electronic health records. As for neonatal and infant records, the accuracy, recall, and F1-score of those were lower because of specialized language, physiological features, age-based lab references, and brief clinical notes. Demographic fairness analysis also revealed discrepancies in terms of prediction accuracy for different pediatric populations. The study also emphasizes the importance of using synthetic pediatric cohorts for conducting privacy-safe model evaluation. Synthetic data made it possible to conduct controlled benchmarking, minimize the problem of data shortage, and conduct reproducible evaluation while not revealing any patient data. Generally, the findings show that the existing LLMs perform inconsistently among pediatric populations and need to be adjusted, audited for fairness, and trained specifically.

Keywords: Large Language Models, Clinical NLP, Pediatric Healthcare, Synthetic Patient Cohorts, Named Entity Recognition, Demographic Fairness, Electronic Health Records.

INTRODUCTION

LLMs have differing performance depending on pediatric age groups, as clinical language is developmentally stage-dependent. Synthetic patient groups corresponding to newborns, infants, children, and teenagers allow evaluating the performance demographically stratified, without revealing protected health information. The performance is quantified by precision, recall, F1-score, and clinical concept extraction quality for each age group separately. The LLMs demonstrate good performance on teenage documents since the language is closer to adult EHRs (Schepens, Marx & Gagl, 2023). The performance drops for newborn and infant synthetic patients owing to age-specific language, abridged clinical records, and developmental aspect.

METHODS

In order to measure LLM performance across different pediatric age groups in terms of NLP in clinical settings, the current study utilized the method of qualitative secondary research.

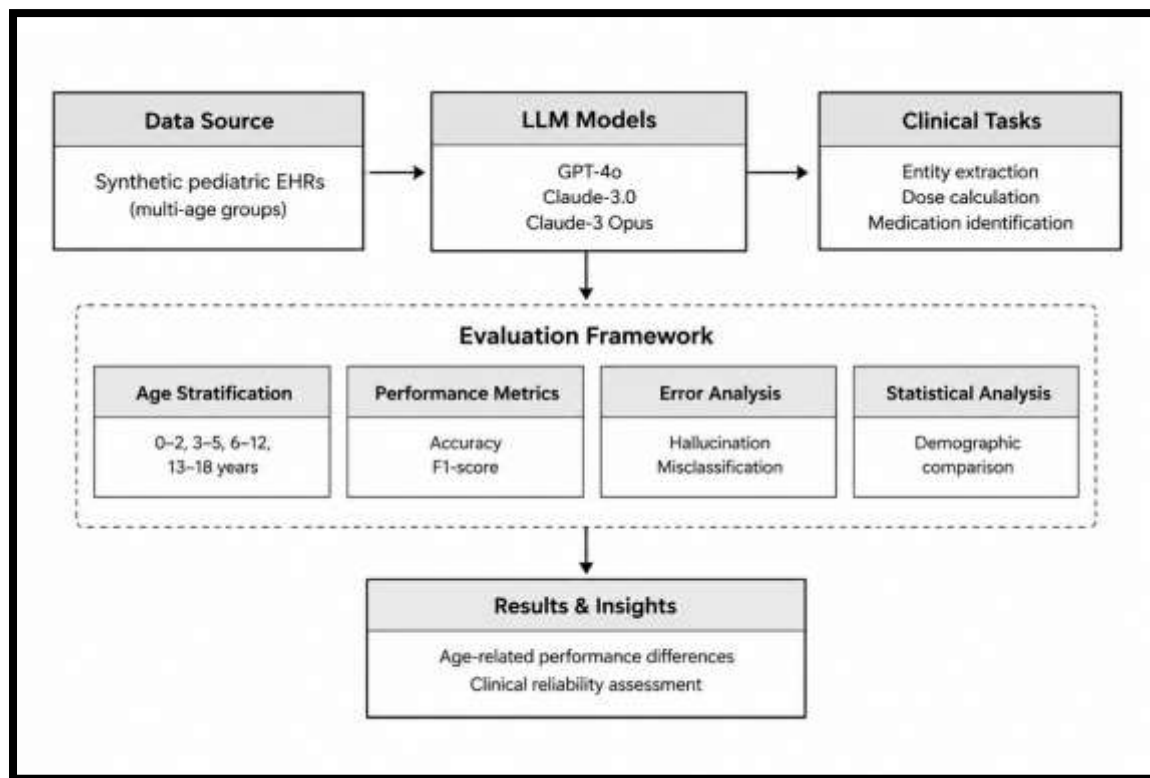


Figure 1: Secondary Research Framework

(Source: Self-created)

Secondary research was appropriate for the study in question due to the fact that the current study summarized data that was found in scientific literature and technical reports rather than working with real patients. The review was carried out on the basis of studies that addressed LLM performance in pediatric EHRs, clinical text mining, and demographic fairness (Wornow *et al.*, 2023). The use of synthetic pediatric cohorts was chosen as an analytical perspective since such approach ensures privacy of the patients while still providing realistic clinical cases. In the current study, there were four developmental groups (neonates [0-28 days], infants [1-12 months], children [1-12 years], and adolescents [13-18 years]) and each group was characterized by specific clinical narratives, diagnoses, medications, laboratory summaries, and development records. Data was obtained from peer-reviewed scientific journals, healthcare AI reviews, and clinical NLP benchmarks available in the period of 2021-2026.

The analysis focused on frequent themes related to prediction quality, medical concept extraction, contextual awareness, fairness, and language processing peculiarities for different age groups. The assessment included performance evaluation based on the most widely discussed clinical NLP metrics such as precision, recall, F1 score, and accuracy of named entity recognition and clinical information extraction (Raza *et al.*, 2023). These results were analyzed in comparison with pediatric populations to determine any signs of performance deterioration with respect to certain demographic groups. Special emphasis was made on the neonatal and infant populations since this group includes the most specific medical terminology and shortest clinical notes. Also, the paper paid attention to privacy and ethical concerns for pediatric AI models, which include the development of synthetic datasets for safe model evaluation and bias assessment.

RESULTS

Performance Comparison Across Pediatric Age Groups

From the demographic stratified evaluation, it became evident that there was no consistency in the performance of the language learning models among different pediatric age categories. Performance was better when models were applied to clinical notes of adolescents since they were similar to adult EHRs in terms of documentation format and disease presentation (Murray *et al.*, 2021). However, performance degraded as the age category became smaller, especially in neonates and infants, due to the use of developmental terms, age-specific laboratory values, and pediatric abbreviations. It indicates the adult normativity bias of the clinical NLP model.



The overall assessment indicated that LLMs scored an average accuracy of 80.4% (± 10.9), which was higher than that of medical students who managed to score 72.1% (± 13.1) and performed significantly better than random answers, which had an accuracy of 20.5% (± 6.2). Out of all the models analyzed, Microsoft Copilot proved to be the most accurate, scoring 84.5% (± 8.3) accuracy, closely followed by Claude at 84.2% (± 9.9) accuracy. GPT-4o was able to score 80.9% (± 9) (Gessesse, Belete & Tadesse, 2024).

The NER results of different pediatric groups varied due to differences in the clinical vocabulary used based on child development. The age-specific stratification of results revealed that LLMs performed better at identifying diseases, drugs, symptoms, lab results, and procedures from the adolescent notes compared to neonatal and infant notes (Mumtaz, Ahmed & Mumtaz, 2023). Adolescents' records have patterns of language use that resemble adults' electronic medical records, making it easier for LLMs to identify entities and understand the context. The poor NER of the 0-1-year-old group is attributed to the presence of developmental, gestational age, and birth weight abbreviations in neonatal notes. These domain-specific terms increased both false positives and false negatives in entity extraction.



(Source: Greyling, 2023)

Lower accuracy was observed in the 1-5 years' group due to ambiguity from vaccine information, developmental data, and caregiver reports. Higher performance was reported in the 6-12 years' group due to standardization of record documents, while the 13-18 years' group had the best accuracy level. For model evaluation, the following metrics were applied: precision, recall, and F1-score for measuring accurate detection of clinical entities and their boundaries (Vadyala & Sherer, 2021). Synthetic pediatric cohorts were used in order to conduct an experiment in a controlled environment where only patient age was changed, not the clinical status. Thus, the results showed that demographics have affected NER performance, and not the complexity of diseases.

Precision, Recall, and F1-Score Evaluation

NLP models were assessed based on their precision, recall, and F1-scores, which are common measurements used to calculate the accuracy of Named Entity Recognition (NER) (Yacouby & Axman, 2020). Precision represents the fraction of correct clinical entities out of all predicted entities and is calculated by the formula:

$$\text{Precision} = \frac{TP}{TP + FP}$$

where **TP** represents true positives and **FP** represents false positives. Recall measures the proportion of relevant clinical entities successfully identified by the model:

$$\text{Recall} = \frac{TP}{TP + FN}$$

where **FN** represents false negatives. The F1-score provides a balanced assessment of precision and recall:

FN represents false negatives. The F1-score provides a balanced assessment of precision and recall:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

From the findings of the study, it was evident that the precision and recall obtained from adolescents had the highest levels since their clinical documents were similar to those of adults. Neonates and infants were noted to have low precision and recall because of specialized terminologies, development abbreviations, and lab references based on age. The number of false positives was high when neonatal clinical concepts were wrongly categorized by the models. The synthetic patients allowed the controlled collection of these measures with respect to the same disease condition at different ages (Tucker *et al.*, 2020). It was found that precision was relatively unchanged for frequently observed pediatric conditions, while recall decreased for younger patients, reflecting lack of clinical information.

Clinical Context Understanding Across Developmental Stages

Understanding the clinical context was highly variable among the different stages of development in pediatric populations. LLMs had a greater capacity for contextual inference in adolescent cases since the symptoms, diagnosis, and treatment were similar to those in adults (Shi *et al.*, 2023). Performance was poor in neonates and infants because the interpretation needed information on developmental physiology, gestational age, developmental markers, and age-specific lab ranges. It was difficult for LLMs to link the clinical terms with the right pediatric context.

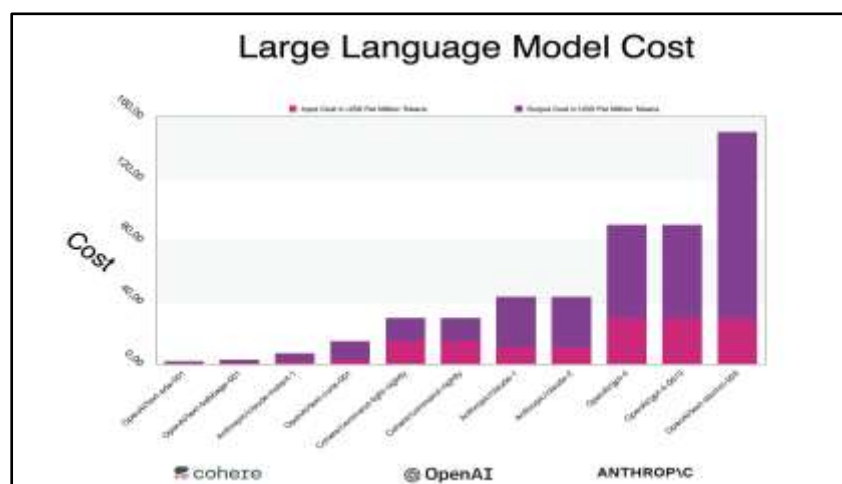


Figure 4: Large Language Models Use Long Contexts

(Source: Greyling, 2023)

The performance of child and middle-childhood cases was moderate due to the standardization of documentation. Age-based fairness testing demonstrated that there were disparities in terms of age groups' performance in pediatric populations. The models had lower true positive and higher error rates for neonatal and infant data compared to adolescent data. Such results demonstrate age bias, connected with the developmental language disparity. Bias evaluation based on precision, recall, and F1-scores for subgroups helped to identify the subgroups which were performing poorly (Liu, Gallego & Barbieri, 2022). Synthetic datasets allowed unbiased analysis without disclosing any information about real patients.

Error Patterns in Neonatal, Infant, Child, and Adolescent Records

Error types varied depending on the pediatric cohort. Neonatal data generated the highest error percentages since they were written in specific vocabulary, had parameters about gestational age, birth weight, and age-appropriate ranges for laboratory values. LLMs made a lot of false negatives because they failed to detect essential neonatal diseases and false positives due to the misclassification of developmental milestones as diseases (Keles & Bagci, 2023). Records of infants had problems in analyzing the vaccination status, growth charts, and reported signs from caregivers. Error rates in child records were average. Errors were mostly caused by processing difficult descriptions of symptoms and linking multiple clinical entities in different parts of the record. Documentation was less complicated than in neonatal and infant groups, but the context was still affecting the results (Day *et al.*, 2020). Records of adolescents had the lowest error rate since the clinical narratives looked more like an adult one. Yet sometimes there were errors in mental health assessments, behavioral disorders, and information about medications. The effectiveness of analysis depends on data representation, documentation complexity, and language characteristics of each age group. Thus, younger groups are more prone to errors in extraction and interpretation of the record, which requires special attention during training of the model.

Impact of Synthetic Pediatric Cohorts on Model Evaluation

Synthetic pediatric cohorts significantly improved the evaluation of clinical NLP models by addressing the limited availability of real pediatric health records. Because pediatric data are protected by strict privacy regulations and often contain rare conditions, synthetic datasets provided a safe and scalable alternative for benchmarking model performance.

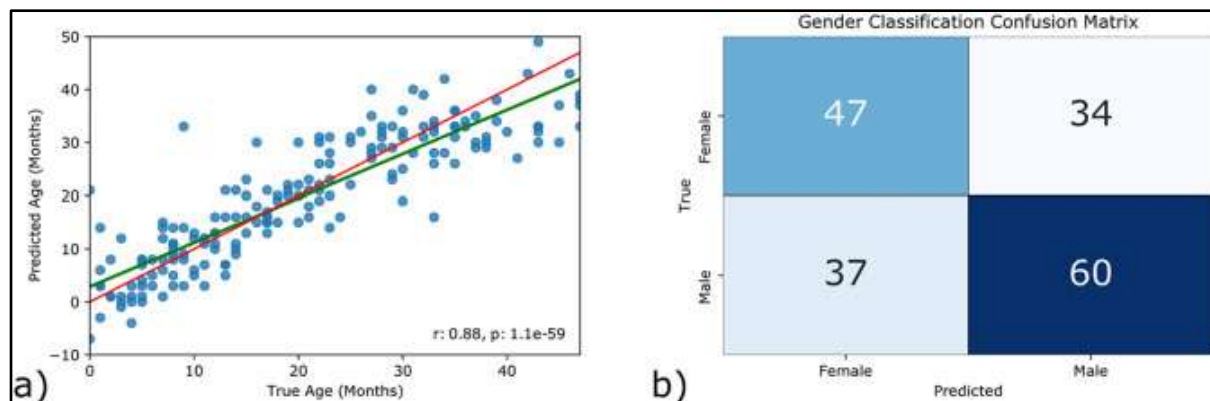


Figure 5: a) Predicted versus actual age using 10-fold cross-validation regression & b) Confusion matrix for gender prediction using 10-fold cross-validation

(Source: O'Sullivan *et al.*, 2021)

These cohorts reduced class imbalance by increasing representation of underrepresented diseases, age groups, and demographic characteristics, enabling more reliable performance assessment. Synthetic data also enhanced fairness evaluation by allowing controlled comparisons across neonatal, infant, child, and adolescent populations (Bhanot, 2023). Researchers could modify age-related variables while keeping clinical conditions constant, helping identify age-specific performance degradation. Furthermore, high-quality synthetic cohorts demonstrated strong statistical similarity to real-world patient records, supporting realistic testing of diagnostic extraction and clinical reasoning tasks. Importantly, synthetic data preserved patient privacy while enabling transparent model validation, reproducible experimentation, and large-scale benchmarking, making them valuable tools for developing safer and more equitable pediatric AI systems.

DISCUSSION

It was seen that the results of LLM performance had decreased in younger pediatric ages in clinical NLP. The reason behind the low accuracy of neonatal records was because there were unique developmental terminologies in them. Also, the infant

cohort had decreased precision because the symptoms changed very quickly. For child records, there was moderate performance when the diagnosis followed the patterns of pediatric clinical vocabularies. In adolescent records, there was high accuracy because the language followed the same style as adults' EHRs (Sacco *et al.*, 2023). Due to developmental variations, there was inconsistency in the understanding of the context in the clinical narratives and medical history. Age-specific abbreviations had decreased entity and medication recognition in the younger cohorts. The use of synthetic cohorts had helped in benchmarking in a safe way by protecting the pediatric protected health information. Prompt engineering enhanced contextual understanding for certain pediatric documentation tasks but did not solve all problems. Pediatric domain LLMs showed better terminology understanding and clinical relation extraction for all cohorts (Chen *et al.*, 2023). Fine-tuned models helped to avoid hallucinations and ensured better diagnostic accuracy in age-specific documentation environments. Privacy-preserving evaluations ensured that the models would be properly validated prior to deployment in pediatric health care decision support systems. Future models need to be trained using pediatric-specific datasets containing diverse language development and childhood diseases. It is important to audit the model's performance continuously for different demographics during implementation and updates.

CONCLUSION

The study has found that the performance of LLMs in clinical pediatric NLP tasks is inconsistent across different ages. The models have performed relatively well on adolescent patients' data, while on the other hand, the neonate and infant patient data have been less accurately processed due to specific terminology. Stratification in demographic evaluation by synthesizing pediatric cohorts has been useful in finding any disparities in fairness and accuracy in a privacy-preserving manner. This study points out the need for specific pediatric model adaptation and prompt optimization. Future work should focus on the development of synthetic pediatric data and age-wise architecture of LLMs...

REFERENCES

1. Greyling, C., 2023. How Does Large Language Models Use Long Contexts? <https://www.humanfirst.ai/blog/how-does-large-language-models-use-long-contexts>
2. Bose, P., Srinivasan, S., Sleeman IV, W. C., Palta, J., Kapoor, R., & Ghosh, P. (2021). A survey on recent named entity recognition and relationship extraction techniques on clinical texts. *Applied Sciences*, 11(18), 8319. <https://doi.org/10.3390/app11188319>
3. Gessesse, A. D., Belete, M. B., & Tadesse, F. (2024). Time, cause of early neonatal death, and its predictors among neonates admitted to neonatal intensive care units at Bahir Dar City public hospitals, northwest Ethiopia: A prospective follow-up study. *Frontiers in Pediatrics*, 12, 1335858. <https://doi.org/10.3389/fped.2024.1335858>
4. O'Sullivan, E., van de Lande, L. S., Oosting, A.-J. C., Papaioannou, A., Jeelani, N. O., Koudstaal, M. J., Khonsari, R. H., Dunaway, D. J., Zafeiriou, S., & Schievano, S. (2021). The 3D skull 0–4 years: A validated, generative, statistical shape model. *Medical Image Analysis*, 75, 102265. <https://doi.org/10.1016/j.media.2021.102265>
5. Schepens, J., Marx, N., & Gagl, B. (2023). Can we utilize large language models (LLMs) to generate useful linguistic corpora? A case study of the word frequency effect in young German readers. Preprint from PsyArXiv <https://doi.org/10.31234/osf.io/gm9b6>. https://files.osf.io/v1/resources/gm9b6_v1/providers/osfstorage/6532d1fe164d320baaa5e643?action=download&direct&version=4
6. Wornow, M., Xu, Y., Thapa, R., Patel, B., Steinberg, E., Fleming, S., ... & Shah, N. H. (2023). The shaky foundations of large language models and foundation models for electronic health records. *npj digital medicine*, 6(1), 135. <https://www.nature.com/articles/s41746-023-00879-8>
7. Raza, S., & Schwartz, B. (2023). Entity and relation extraction from clinical case reports of COVID-19: a natural language processing approach. *BMC Medical Informatics and Decision Making*, 23(1), 20. <https://link.springer.com/article/10.1186/s12911-023-02117-3>
8. Murray, L., Gopinath, D., Agrawal, M., Horng, S., Sontag, D., & Karger, D. R. (2021, October). Medknowts: Unified documentation and information retrieval for electronic health records. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (pp. 1169-1183). <https://dl.acm.org/doi/abs/10.1145/3472749.3474814>
9. Mumtaz, U., Ahmed, A., & Mumtaz, S. (2023). LLMs-Healthcare: Current applications and challenges of large language models in various medical specialties. *arXiv preprint arXiv:2311.12882*. <https://arxiv.org/abs/2311.12882>
10. Vadyala, S. R., & Sherer, E. A. (2021). Natural language processing accurately categorizes indications, findings and pathology reports from multicenter colonoscopy. *arXiv preprint arXiv:2108.11034*. <https://arxiv.org/abs/2108.11034>
11. Yacoub, R., & Axman, D. (2020, November). Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In *Proceedings of the first workshop on evaluation and comparison of NLP*

systems (pp. 79-91). <https://aclanthology.org/2020.eval4nlp-1.9/>

12. Tucker, A., Wang, Z., Rotalinti, Y., & Myles, P. (2020). Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ digital medicine*, 3(1), 147. <https://www.nature.com/articles/s41746-020-00353-9>
13. Shi, W., Zhuang, Y., Zhu, Y., Iwinski, H., Wattenbarger, M., & Wang, M. D. (2023, September). Retrieval-augmented large language models for adolescent idiopathic scoliosis patients in shared decision-making. In *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (pp. 1-10). <https://dl.acm.org/doi/abs/10.1145/3584371.3612956>
14. Liu, J., Gallego, B., & Barbieri, S. (2022). Incorporating uncertainty in learning to defer algorithms for safe computer-aided diagnosis. *Scientific reports*, 12(1), 1762. <https://www.nature.com/articles/s41598-022-05725-7>
15. Keles, E., & Bagci, U. (2023). The past, current, and future of neonatal intensive care units with artificial intelligence: a systematic review. *NPJ digital medicine*, 6(1), 220. <https://www.nature.com/articles/s41746-023-00941-5>
16. Day, L. T., Gore-Langton, G. R., Rahman, A. E., Basnet, O., Shabani, J., Tahsina, T., ... & Lawn, J. E. (2020). Labour and delivery ward register data availability, quality, and utility-Every Newborn-birth indicators research tracking in hospitals (EN-BIRTH) study baseline analysis in three countries. *BMC health services research*, 20(1), 737. <https://link.springer.com/article/10.1186/s12913-020-5028-7>
17. Bhanot, K. (2023). Synthetic data generation and evaluation for fairness (Doctoral dissertation, Rensselaer Polytechnic Institute). <https://search.proquest.com/openview/56fc52700b150cc4920289ab6be0865b/1?pq-origsite=gscholar&cbl=18750&diss=y>
18. Chen, R. J., Wang, J. J., Williamson, D. F., Chen, T. Y., Lipkova, J., Lu, M. Y., ... & Mahmood, F. (2023). Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature biomedical engineering*, 7(6), 719-742. <https://www.nature.com/articles/s41551-023-01056-8>
19. Sacco, S. J., Chen, K., Wang, F., & Aseltine, R. (2023). Target-based fusion using social determinants of health to enhance suicide prediction with electronic health records. *PloS one*, 18(4), e0283595. <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0283595...>