UC San Diego
SCHOOL OF MEDICINE

Hannah Carter, PhD
Assistant Professor, Medicine
University of California, San Diego
9500 Gilman Drive, MC 0688
La Jolla, CA 92093-0688
Email: hkcarter@health.ucsd.edu
Phone: +1 (858) 822-4706

03/05/2021
Re: Revised submission to Database

Dear Dr. Landsman,

Thank you for coordinating the peer review of our manuscript. In this manuscript, we address the state of metadata quality in the Sequence Read Archive (SRA) hosted by NCBI. We highlight several issues in the annotations of samples deposited in the SRA, that have similarly been reported for other large data repositories and propose a novel deep learning-based approach for metadata curation of SRA samples, utilizing Named Entity Recognition (NER) to improve the plurality and completeness of the metadata landscape.

We have now revised our manuscript according to the guidance of the three reviewers and believe doing so has strengthened it considerably. These changes include providing more details on reproducibility and how parameter values were selected, as well as additional context and clarification of where our work fits within the field. Specifically, we show that multiple trials of our workflow produce very similar results and have added Supplementary Table 3 to highlight this. We also justified the cosine similarity thresholds used (Supplementary Table 4) and the decision to include unigrams as part of the ground truth in our predictions (Supplementary Figure 4). We have also expanded our discussion to better address possible future directions of this work.

We are attaching a point-by-point response to reviewer comments, detailing how we have updated our manuscript. We look forward to next steps towards publishing our work with Database.

Sincerely,

Hannah Carter