香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

UNIVERSITY OF TORONTO

CReATe
Fertility Centre

中信湘雅生殖与遗传专科医院
REPRODUCTIVE & GENETIC HOSPITAL OF CITIC-XIANGYA

深圳市罗湖医院集团罗湖区人民医院
Shenzhen Luohu Hospital Group Luohu People's Hospital
深圳大学第三附属医院
The third Affiliated Hospital of Shenzhen University

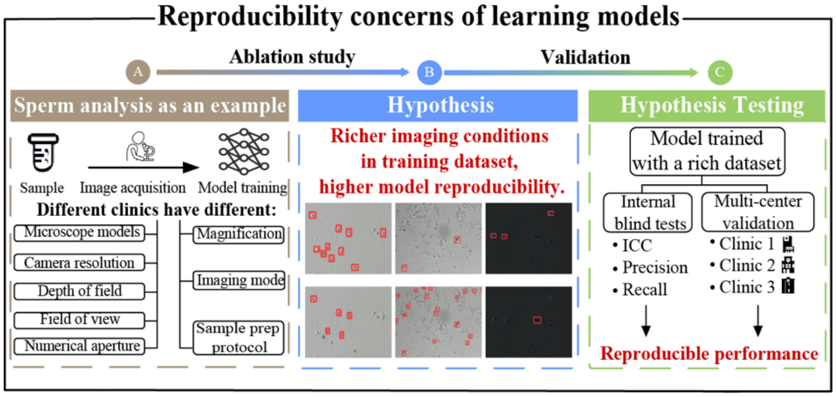
January 24th, 2024

Dear Editor,

We are writing to submit a manuscript titled “Testing the reproducibility and effectiveness of deep learning models among clinics - sperm detection as a pilot study”, which we would like to have considered for publication in *npj Digital Medicine.* This work has not been published nor is under consideration for publication elsewhere.

Recently numerous deep learning models have been developed to facilitate the diagnosis and treatment of infertility, which is a global health issue affecting one in six couples worldwide. Although it is commonly believed that a “richer” training dataset is necessary for “better” models, most of existing research was devoted to improving the accuracy of the models. However, model reproducibility has been largely ignored. Different clinics use different imaging conditions, raising the concern over whether the reported accuracy in one clinic could be reproduced in another clinic (see Figure below). Unfortunately, it remains poorly understood how richness in the training dataset impacts model reproducibility.

Here we fill this gap by quantitatively investigating how different imaging conditions in different clinics affect model reproducibility. We used sperm detection as an example, which is the most widely used deep learning technique for infertility diagnosis and treatment, and tested the reproducibility of state-of-the-art sperm detection models among different clinics. Our contributions are three-fold:



(1) Ablation studies quantitatively revealed how model precision (false-positive detection) and recall (missed detection) were affected by image magnification, imaging mode, and sample preprocessing approaches.

(2) Based on results from the ablation studies, we hypothesize that a training dataset containing images collected under diverse imaging conditions could improve model reproducibility for clinical deployment. This hypothesis was first tested via internal blind tests where the same sample was repeatedly measured by the same model but under different imaging conditions to quantify intraclass correlation coefficient. **The hypothesis was further validated by external multi-center validation in different clinical applications** (male infertility diagnosis, and calculating sperm concentration for *in vitro* fertilization treatment).

(3) Our results also provide clinical users with new guidelines for evaluating deep learning models. Model accuracy or AUC is not the sole metric. Before deploying a model, practitioners should pay more attention to the training dataset and judge whether it contains the imaging condition of their own clinics’.

Thank you for considering our manuscript. We hope you agree that it meets the high standards required for publication in *npj Digital Medicine*, and look forward to receiving review comments.

Sincerely,

Zhuoran Zhang (on behalf of all co-authors)
The Chinese University of Hong Kong, Shenzhen