KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

Citation/Reference	Van Calster, B., Timmerman, S., Geysels, A., Verbakel, J.Y., Froyman, W. with Froyman, W. (corresp. author) (2022). A deep-learning-enabled diagnosis of ovarian cancer. *LANCET DIGITAL HEALTH*, *4* (9), E630-E630.
Archived version	Author manuscript: the content is identical to the content of the published paper.
Published version	doi: 10.1016/S2589-7500(22)00130-3
Journal homepage	https://www.thelancet.com/journals/landig/home
Author contact	axel.geysels@esat.kuleuven.be
IR	https://lirias.kuleuven.be/retrieve/674544

(article begins on next page)

A deep-learning-enabled diagnosis of ovarian cancer

We read with interest the results of the retrospective multicentre study on deep learning-enabled diagnosis of ovarian cancer by Yue Gao and colleagues,[1] which is the largest study conducted on this topic to our knowledge. Nevertheless, we have several methodological concerns that warrant caution regarding the study's conclusions.

First, although the goal of the deep convolutional neural network was to discriminate between benign and ovarian masses, the training dataset included healthy controls. These individuals do not belong to the target population (ie, patients with ovarian masses) and hence distort predictions for those who do.[2] Additionally, the number of healthy controls and their inclusion and exclusion criteria were not described.

Second, the data curation process was not clearly described. Without adequate care, the model might leverage clinician-based information that is inherently present in retrospectively collected JPEG images, such as annotations, calipers, or Doppler information. If so, model performance might be overoptimistic and generalisability might be affected.[3] Interpretation methods, such as saliency maps, could give insight into this process and explain some of the black-box decisions that are made by the deep convolutional neural network.

Third, the conclusion that the model outperforms the average diagnostic level of radiologists might be softened. The reported performance of the radiologists in the study is low compared with the performance that is described elsewhere in the scientific literature.[4]

Fourth, the calibration performance of the deep convolutional neural network was poor with underestimated risks of malignancy, which could lead to undertreatment when used in practice.

Gao and colleagues[1] refer to ultrasound-based risk prediction algorithms, such as the ADNEX logistic regression model, as "effective methods to discriminate between benign and malignant adnexal masses". Indeed, several external validation studies by investigators have reported AUCs above 0·9 with decent calibration performance (ie, with calibration curves close to the ideal diagonal line).[5] In conclusion, we are concerned about the potential value and face validity of this model.

We declare no competing interests.

Copyright © 2022 The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY-NC-ND 4.0 license.

*Ben Van Calster, Stefan Timmerman, Axel Geysels, Jan Y Verbakel, *Wouter Froyman*

wouter.froyman@uzleuven.be

Department of Development and Regeneration (BVC, ST, WF), EPI-Centre, Department of Public Health and Primary Care (BVC, JYV), and Department of Electrical Engineering, STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics (AG), KU Leuven, Leuven, Belgium; Department of Biomedical Data Sciences, Leiden University Medical Centre, Leiden, Netherlands (BVC); Department of Obstetrics and Gynaecology, University Hospitals Leuven, Leuven 3000, Belgium (ST, WF); Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK (JYV)

1 Gao Y, Zeng S, Xu X, et al. Deep learning-enabled pelvic ultrasound images for accurate diagnosis of ovarian cancer in China: a retrospective, multicentre, diagnostic study. *Lancet Digit Health* 2022; **4:** e179–87.

2 Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999; **282:** 1061–66.

3 Badgeley MA, Zech JR, Oakden-Rayner L, et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digit Med* 2019; **2:** 31.

4 Meys EM, Kaijser J, Kruitwagen RF, et al. Subjective assessment versus ultrasound models to diagnose ovarian cancer: a systematic review and meta-analysis. *Eur J Cancer* 2016; **58:** 17–29.

5 Van Calster B, Valentin L, Froyman W, et al. Validation of models to diagnose ovarian cancer in patients managed surgically or conservatively: multicentre cohort study. *BMJ* 2020; **370:** m2614.