npj | digital medicine

www.nature.com/npjdigitalmed

REVIEW ARTICLE OPEN

Machine learning for medical imaging: methodological failures and recommendations for the future

Gaël Varoquaux[1,2,3]✉ and Veronika Cheplygina[4]✉

Research in computer analysis of medical images bears many promises to improve patients' health. However, a number of systematic challenges are slowing down the progress of the field, from limitations of the data, such as biases, to research incentives, such as optimizing for publication. In this paper we review roadblocks to developing and assessing methods. Building our analysis on evidence from the literature and data challenges, we show that at every step, potential biases can creep in. On a positive note, we also discuss on-going efforts to counteract these problems. Finally we provide recommendations on how to further address these problems in the future.

npj Digital Medicine (2022)5:48 ; https://doi.org/10.1038/s41746-022-00592-y

INTRODUCTION

Machine learning, the cornerstone of today's artificial intelligence (AI) revolution, brings new promises to clinical practice with medical images[1–3]. For example, to diagnose various conditions from medical images, machine learning has been shown to perform on par with medical experts[4]. Software applications are starting to be certified for clinical use[5,6]. Machine learning may be the key to realizing the vision of AI in medicine sketched several decades ago[7].

The stakes are high, and there is a staggering amount of research on machine learning for medical images. But this growth does not inherently lead to clinical progress. The higher volume of research could be aligned with the academic incentives rather than the needs of clinicians and patients. For example, there can be an oversupply of papers showing state-of-the-art performance on benchmark data, but no practical improvement for the clinical problem. On the topic of machine learning for COVID, Robert et al.[8] reviewed 62 published studies, but found none with potential for clinical use.

In this paper, we explore avenues to improve clinical impact of machine learning in medical imaging. After sketching the situation, documenting uneven progress in Section It's not all about larger datasets, we study a number of failures frequent in medical imaging papers, at different steps of the "publishing lifecycle": what data to use (Section Data, an imperfect window on the clinic), what methods to use and how to evaluate them (Section Evaluations that miss the target), and how to publish the results (Section Publishing, distorted incentives). In each section, we first discuss the problems, supported with evidence from previous research as well as our own analyses of recent papers. We then discuss a number of steps to improve the situation, sometimes borrowed from related communities. We hope that these ideas will help shape research practices that are even more effective at addressing real-world medical challenges.

IT'S NOT ALL ABOUT LARGER DATASETS

The availability of large labeled datasets has enabled solving difficult machine learning problems, such as natural image recognition in computer vision, where datasets can contain millions of images. As a result, there is widespread hope that similar progress will happen in medical applications, algorithm research should eventually solve a clinical problem posed as discrimination task. However, medical datasets are typically smaller, on the order of hundreds or thousands:[9] share a list of sixteen "large open source medical imaging datasets", with sizes ranging from 267 to 65,000 subjects. Note that in medical imaging we refer to the number of subjects, but a subject may have multiple images, for example, taken at different points in time. For simplicity here we assume a diagnosis task with one image/scan per subject.

Few clinical questions come as well-posed discrimination tasks that can be naturally framed as machine-learning tasks. But, even for these, larger datasets have to date not lead to the progress hoped for. One example is that of early diagnosis of Alzheimer's disease (AD), which is a growing health burden due to the aging population. Early diagnosis would open the door to early-stage interventions, most likely to be effective. Substantial efforts have acquired large brain-imaging cohorts of aging individuals at risk of developing AD, on which early biomarkers can be developed using machine learning[10]. As a result, there have been steady increases in the typical sample size of studies applying machine learning to develop computer-aided diagnosis of AD, or its predecessor, mild cognitive impairment. This growth is clearly visible in publications, as on Fig. 1a, a meta-analysis compiling 478 studies from 6 systematic reviews[4,11–15].

However, the increase in data size (with the largest datasets containing over a thousand subjects) did not come with better diagnostic accuracy, in particular for the most clinically relevant question, distinguishing pathological versus stable evolution for patients with symptoms of prodromal Alzheimer's (Fig. 1b). Rather, studies with larger sample sizes tend to report worse prediction accuracy. This is worrisome, as these larger studies are closer to real-life settings. On the other hand, research efforts across time did lead to improvements even on large, heterogeneous cohorts (Fig. 1c), as studies published later show improvements for large sample sizes (statistical analysis in Supplementary Information). Current medical-imaging datasets are much smaller than those that brought breakthroughs in computer vision. Although a one-

[1]INRIA, Versailles, France. [2]McGill University, Montreal, Canada. [3]Mila, Montreal, Canada. [4]IT University of Copenhagen, Copenhagen, Denmark. ✉email: gael.varoquaux@inria.fr; vech@itu.dk

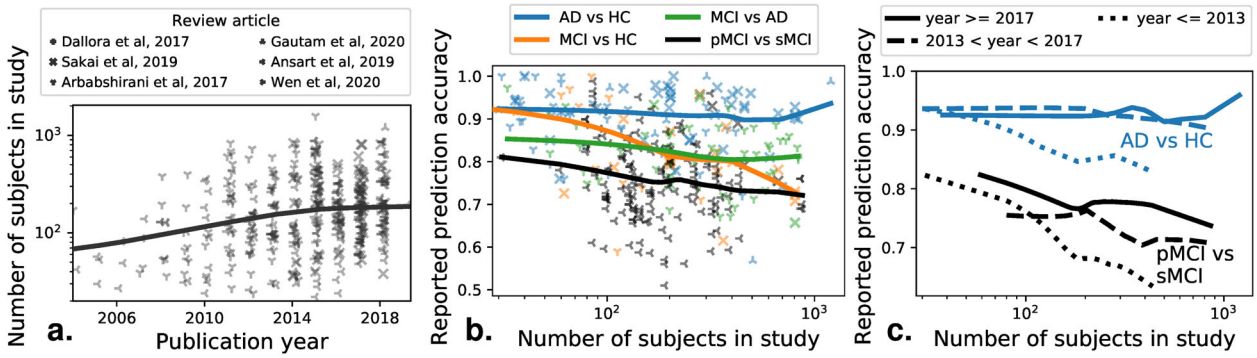

Fig. 1 Larger brain-imaging datasets are not enough for better machine-learning diagnosis of Alzheimer's. A meta-analysis across 6 review papers, covering more than 500 individual publications. The machine-learning problem is typically formulated as distinguishing various related clinical conditions, Alzheimer's Disease (AD), Healthy Control (HC), and Mild Cognitive Impairment, which can signal prodromal Alzheimer's . Distinguishing progressive mild cognitive impairment (pMCI) from stable mild cognitive impairment (sMCI) is the most relevant machine-learning task from the clinical standpoint. **a** Reported sample size as a function of the publication year of a study. **b** Reported prediction accuracy as a function of the number of subjects in a study. **c** Same plot distinguishing studies published in different years.

to-one comparison of sizes cannot be made, as computer vision datasets have many classes with high variation (compared to few classes with less variation in medical imaging), reaching better generalization in medical imaging may require assembling significantly larger datasets, while avoiding biases created by opportunistic data collection, as described below.

DATA, AN IMPERFECT WINDOW ON THE CLINIC

Datasets may be biased: reflect an application only partly

Available datasets only partially reflect the clinical situation for a particular medical condition, leading to dataset bias[16]. As an example, a dataset collected as part of a population study might have different characteristics that people who are referred to the hospital for treatment (higher incidence of a disease). As the researcher may be unaware of the corresponding dataset bias is can lead to important that shortcomings of the study. Dataset bias occurs when the data used to build the decision model (the training data), has a different distribution than the data on which it should be applied[17] (the test data). To assess clinically-relevant predictions, the test data must match the actual target population, rather than be a random subset of the same data pool as the train data, the common practice in machine-learning studies. With such a mismatch, algorithms which score high in benchmarks can perform poorly in real world scenarios[18]. In medical imaging, dataset bias has been demonstrated in chest X-rays[19–21], retinal imaging[22], brain imaging[23,24], histopathology[25], or dermatology[26]. Such biases are revealed by training and testing a model across datasets from different sources, and observing a performance drop across sources.

There are many potential sources of dataset bias in medical imaging, introduced at different phases of the modeling process[27]. First, a cohort may not appropriately represent the range of possible patients and symptoms, a bias sometimes called *spectrum bias*[28]. A detrimental consequence is that model performance can be overestimated for different groups, for example between male and female individuals[21,26]. Yet medical imaging publications do not always report the demographics of the data.

Imaging devices or procedures may lead to specific measurement biases. A bias particularly harmful to clinically relevant automated diagnosis is when the data capture medical interventions. For instance, on chest X-ray datasets, images for the "pneumothorax" condition sometimes show a chest drain, which is a treatment for this condition, and which would not yet be present before diagnosis[29]. Similar spurious correlations can appear in skin lesion images due to markings placed by dermatologists next to the lesions[30].

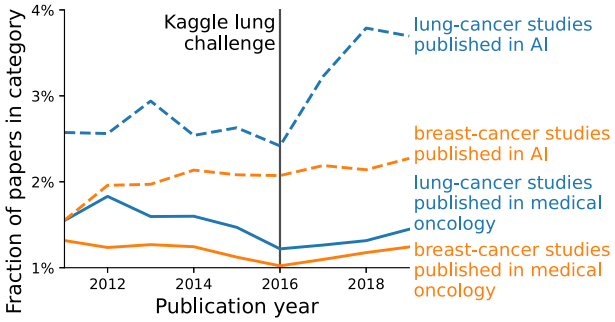

Fig. 2 Differences between relative popularity of applications. We show the percentage of papers on lung cancer (in blue) vs breast cancer (in red), relative to all papers within two fields: medical oncology (solid line) and AI (dotted line). Details on how the papers are selected are given in the Supplementary Information). The percentages are relatively constant, except lung cancer in AI, which shows an increase after 2016.

Labeling errors can also introduce biases. Expert human annotators may have systematic biases in the way they assign different labels[31], and it is seldom possible to compensate with multiple annotators. Using automatic methods to extract labels from patient reports can also lead to systematic errors[32]. For example, a report on a follow-up scan that does not mention previously-known findings, can lead to an incorrect "negative" labels.

Dataset availability distorts research

The availability of datasets can influence which applications are studied more extensively. A striking example can be seen in two applications of oncology: detecting lung nodules, and detecting breast tumors in radiological images. Lung datasets are widely available on Kaggle or grand-challenge.org, contrasted with (to our knowledge) only one challenge focusing on mammograms. We look at the popularity of these topics, here defined by the fraction of papers focusing on lung or breast imaging, either in literature on general medical oncology, or literature on AI. In medical oncology this fraction is relatively constant across time for both lung and breast imaging, but in the AI literature lung imaging publications show a substantial increase in 2016 (Fig. 2, methodological details in Supplementary Information). We suspect that the Kaggle lung challenges published around that time

contributed to this disproportional increase. A similar point on dataset trends has been made throughout the history of machine learning in general[33].

Let us build awareness of data limitations

Addressing such problems arising from the data requires critical thinking about the choice of datasets, at the project level, i.e. which datasets to select for a study or a challenge, and at a broader level, i.e. which datasets we work on as a community.

At the project level, the choice of the dataset will influence the models trained on the data, and the conclusions we can draw from the results. An important step is using datasets from multiple sources, or creating robust datasets from the start when feasible[9]. However, existing datasets can still be critically evaluated for dataset bias[34], hidden subgroups of patients[29], or mislabeled instances[35]. A checklist for such evaluation on computer vision datasets is presented in Zendel et al.[18]. When problems are discovered, relabeling a subset of the data can be a worthwhile investment[36].

At the community level, we should foster understanding of the datasets' limitations. Good documentation of datasets should describe their characteristics and data collection[37]. Distributed models should detail their limitations and the choices made to train them[38].

Meta-analyses which look at evolution of dataset use in different areas are another way to reflect on current research efforts. For example, a survey of crowdsourcing in medical imaging[39] shows a different distribution of applications than surveys focusing on machine learning[1,2]. Contrasting more clinically-oriented venues to more technical venues can reveal opportunities for machine learning research.

EVALUATIONS THAT MISS THE TARGET

Evaluation error is often larger than algorithmic improvements

Research on methods often focuses on outperforming other algorithms on benchmark datasets. But too strong a focus on benchmark performance can lead to *diminishing returns*, where increasingly large efforts achieve smaller and smaller performance gains. Is this also visible in the development of machine learning in medical imaging?

We studied performance improvements in 8 Kaggle medical-imaging challenges, 5 on detection of diagnosis of diseases and 3 on image segmentation (details in Supplementary Information). We use the differences in algorithms performance between the public and private leaderboards (two test sets used in the challenge) to quantify the *evaluation noise* –the spread of performance differences between the public and private test sets–, in Fig. 3. We compare its distribution to the *winner gap*—the difference in performance between the best algorithm, and the "top 10%" algorithm.

Overall, 6 of the 8 challenges are in the diminishing returns category. For 5 challenges—lung cancer, schizophrenia, prostate cancer diagnosis and intracranial hemorrhage detection—the evaluation noise is worse than the winner gap. In other words, the gains made by the top 10% of methods are smaller than the expected noise when evaluating a method.

For another challenge, pneumothorax segmentation, the performance on the private set is worse than on the public set, revealing an overfit larger than the winner gap. Only two challenges (covid 19 abnormality and nerve segmentation) display a winner gap larger than the evaluation noise, meaning that the winning method made substantial improvements compared to the 10% competitor.

Improper evaluation procedures and leakage

Unbiased evaluation of model performance relies on training and testing the models with independent sets of data[40]. However incorrect implementations of this procedure can easily leak information, leading to overoptimistic results. For example some studies classifying ADHD based on brain imaging have engaged in circular analysis[41], performing feature selection on the full dataset, before cross-validation. Another example of leakage arises when repeated measures of an individual are split across train and test set, the algorithm then learning to recognize the individual patient rather than markers of a condition[42].

A related issue, yet more difficult to detect, is what we call "overfitting by observer": even when using cross-validation, overfitting may still occur by the researcher adjusting the method to improve the observed cross-validation performance, which essentially includes the test folds into the validation set of the model. Skocik et al.[43] provide an illustration of this phenomenon by showing how by adjusting the model this way can lead to better-than-random cross-validation performance for randomly generated data. This can explain some of the overfitting visible in challenges (Section Evaluation error is often larger than algorithmic improvements), though with challenges a private test set reveals the overfitting, which is often not the case for published studies. Another recommendation for challenges would be to hold out several datasets (rather than a part of the same dataset), as is for example done in the Decathlon challenge[44].

Metrics that do not reflect what we want

Evaluating models requires choosing a suitable metric. However, our understanding of "suitable" may change over time. For example, an image similarity metric which was widely used to evaluate image registration algorithms, was later shown to be ineffective as scrambled images could lead to high scores[45].

In medical image segmentation, Maier-Hein et al.[46] review 150 challenges and show that the typical metrics used to rank algorithms are sensitive to different variants of the same metric, casting doubt on the objectivity of any individual ranking.

Important metrics may be missing from evaluation. Next to typical classification metrics (sensitivity, specificity, area under the curve), several authors argue for a calibration metric that compares the predicted and observed probabilities[28,47].

Finally, the metrics used may not be synonymous with practical improvement[48,49]. For example, typical metrics in computer vision do not reflect important aspects of image recognition, such as robustness to out-of-distribution examples[49]. Similarly, in medical imaging, improvements in traditional metrics may not necessarily translate to different clinical outcomes, e.g. robustness may be more important than an accurate delineation in a segmentation application.

Incorrectly chosen baselines

Developing new algorithms builds upon comparing these to baselines. However, if these baselines are poorly chosen, the reported improvement may be misleading.

Baselines may not properly account for recent progress, as revealed in machine-learning applications to healthcare[50], but also other applications of machine learning[51–53].

Conversely, one should not forget simple approaches effective for the problem at hand. For example, Wen et al.[14] show that convolutional neural networks do not outperform support vector machines for Alzheimer's disease diagnosis from brain imaging.

Finally, minute implementation details of algorithms may be important and many are not aware of implementation factors[54].

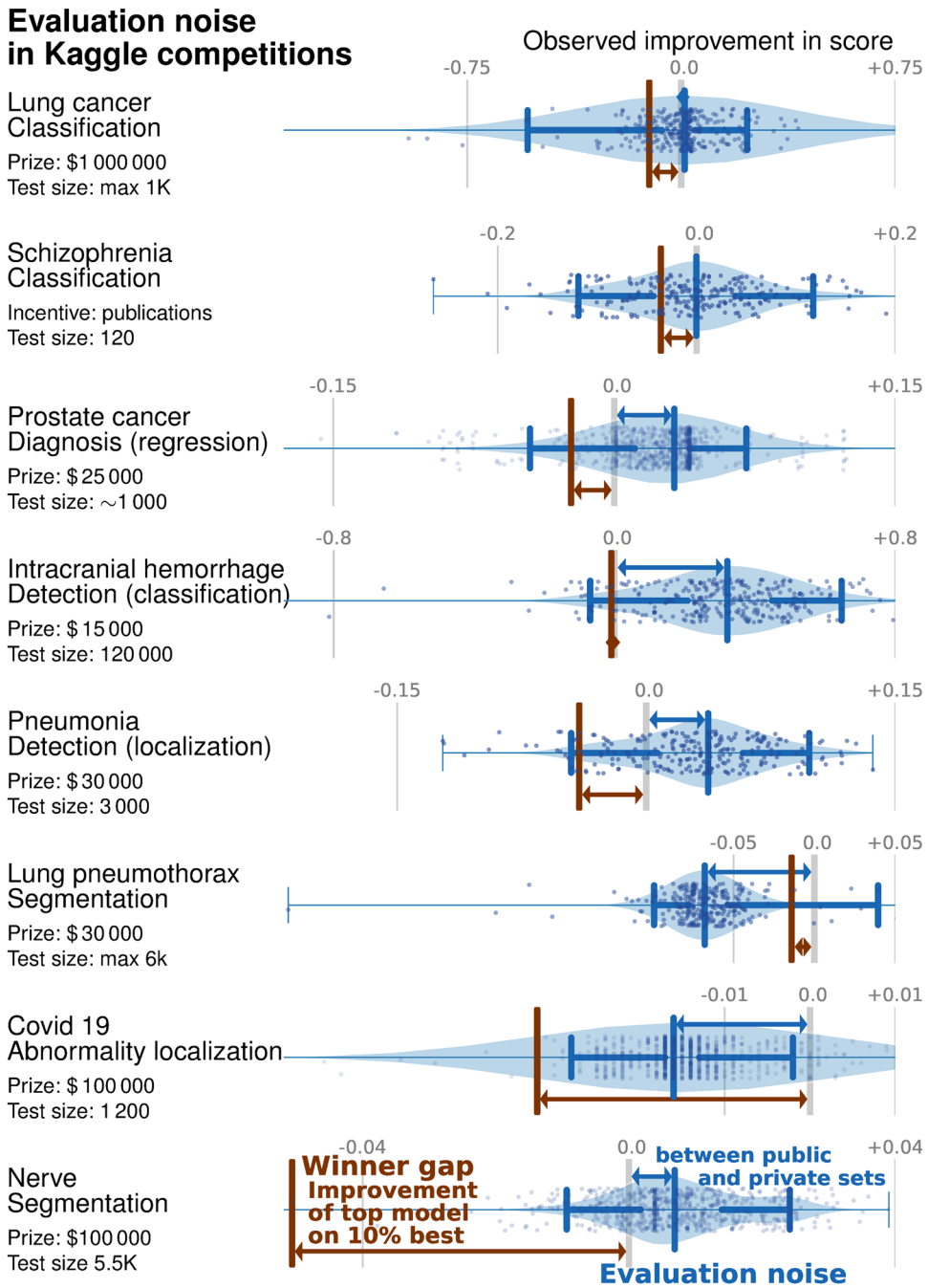

Fig. 3 Kaggle challenges: shifts from public to private set compared to improvement across the top 10% models on eight medical-imaging challenges with significant incentives. The blue violin plot shows the *evaluation noise*—the distribution of differences between public and private leaderboards. A systematic shift between public and private set (positive means that the private leaderboard is better than the public leaderboard) indicates overfitting or dataset bias. The width of this distribution shows how noisy the evaluation is, or how representative the public score is for the private score. The brown bar is the *winner gap*, the improvement between the top-most model (the winner) and the 10% best model. It is interesting to compare this improvement to the shift and width in the difference between the public and private sets: if the winner gap is smaller, the 10% best models reached diminishing returns and did not lead to a actual improvement on new data.

Statistical significance not tested, or misunderstood

Experimental results are by nature noisy: results may depend on which specific samples were used to train the models, the random initializations, small differences in hyper-parameters[55]. However, benchmarking predictive models currently lacks well-adopted statistical good practices to separate out noise from generalizable findings.

A first, well-documented, source of brittleness arises from machine-learning experiments with too small sample sizes[56]. Indeed, testing predictive modeling requires many samples, more than conventional inferential studies, else the measured prediction accuracy may be a distant estimation of real-life performance. Sample sizes are growing, albeit slowly[57]. On a positive note, a meta-analysis of public vs private leaderboards on Kaggle[58]

suggests that overfitting is less of an issue with "large enough" test data (at least several thousands).

Another challenge is that strong validation of a method requires it to be robust to details of the data. Hence validation should go beyond a single dataset, and rather strive for statistical consensus across multiple datasets[59]. Yet, the corresponding statistical procedures require dozens of datasets to establish significance and are seldom used in practice. Rather, medical imaging research often reuses the same datasets across studies, which raises the risk of finding an algorithm that performs well by chance, in an implicit multiple comparison problem[60].

But overall medical imaging research seldom analyzes how likely empirical results are to be due to chance: only 6% of segmentation challenges surveyed[61], and 15% out of 410 popular computer science papers published by ACM used a statistical test[62].

However, null-hypothesis tests are often misinterpreted[63], with two notable challenges: (1) the lack of statistically significant results does not demonstrate the absence of effect, and (2) any trivial effect can be significant given enough data[64,65]. For these reasons, Bouthiellier et al.[66] recommend to replace traditional null-hypothesis testing with *superiority testing*, testing that the improvement is above a given threshold.

Let us redefine evaluation

Higher standards for benchmarking. Good machine-learning benchmarks are difficult. We compile below several recognized best practices for medical machine learning evaluation[28,40,67,68]:

- Safeguarding from data leakage by separating out all test data from the start, before any data transformation.
- A documented way of selecting model hyper-parameters (including architectural parameters for neural networks, the use of additional (unlabeled) dataset or transfer learning[2]), without ever using data from the test set.
- Enough data in the test set to bring statistical power, at least several hundreds samples, ideally thousands or more[9], and confidence intervals on the reported performance metric—see Supplementary Information. In general, more research on appropriate sample sizes for machine learning studies would be helpful.
- Rich data to represent the diversity of patients and disease heterogeneity, ideally multi-institutional data including all relevant patient demographics and disease state, with explicit inclusion criteria; other cohorts with different recruitment go the extra mile to establish external validity[69,70].
- Strong baselines that reflect the state of the art of machine-learning research, but also historical solutions including clinical methodologies not necessarily relying on medical imaging.
- A discussion the variability of the results due to arbitrary choices (random seeds) and data sources with an eye on statistical significance—see Supplementary Information.
- Using different quantitative metrics to capture the different aspects of the clinical problem and relating them to relevant clinical performance metrics. In particular, the potential health benefits from a detection of the outcome of interest should be used to choose the right trade off between false detections and misses[71].
- Adding qualitative accounts and involving groups that will be most affected by the application in the metric design[72].

More than beating the benchmark. Even with proper validation and statistical significance testing, measuring a tiny improvement on a benchmark is seldom useful. Rather, one view is that, beyond rejecting a null, a method should be accepted based on evidence that it brings a sizable improvement upon the existing solutions. This type of criteria is related to *superiority tests* sometimes used in clinical trials[73–75]. These tests are easy to implement in predictive modeling benchmarks, as they amount to comparing the observed improvement to variation of the results due to arbitrary choices such as data sampling or random seeds[55].

Organizing blinded challenges, with a hidden test set, mitigate the winner's curse. But to bring progress, challenges should not only focus on the winner. Instead, more can be learned by comparing the competing methods and analyzing the determinants of success, as well as failure cases.

Evidence-based medicine good practices. A machine-learning algorithm deployed in clinical practice is a health intervention. There is a well-established practice to evaluate the impact of health intervention, building mostly on randomized clinical trials[76]. These require actually modifying patients' treatments and thus should be run only after thorough evaluation on historical data.

A solid trial evaluates a well-chosen measure of patient health outcome, as opposed to predictive performance of an algorithm. Many indirect mechanisms may affect this outcome, including how the full care processes adapts to the computer-aided decision. For instance, a positive consequence of even imperfect predictions may be reallocating human resources to complex cases. But a negative consequence may be over-confidence leading to an increase in diagnostic errors. Cluster randomized trials can account for how modifications at the level of care unit impact the individual patient: care units, rather than individuals are randomly allocated to receive the intervention (the machine learning algorithm)[77]. Often, double blind is impossible: the care provider is aware of which arm of the study is used, the baseline condition or the system evaluated. Providers' expectations can contribute to the success of a treatment, for instance via indirect placebo or nocebo effects[78], making objective evaluation of the health benefits challenging, if these are small.

PUBLISHING, DISTORTED INCENTIVES

No incentive for clarity

The publication process does not create incentives for clarity. Efforts to impress may give rise to unnecessary "mathiness" of papers or suggestive language[79] (such as "human-level performance").

Important details may be omitted, from ablation experiments showing what part of the method drives improvements[79], to reporting how algorithms were evaluated in a challenge [46]. This in turn undermines reproducibility: being able to reproduce the exact results or even draw the same conclusions[80,81].

Optimizing for publication

As researchers our goal should be to solve scientific problems. Yet, the reality of the culture we exist in can distort this objective. Goodhart's law summarizes well the problem: *when a measure becomes a target, it ceases to be a good measure*. As our academic incentive system is based publications, it erodes their scientific content via Goodhart's law.

Methods publication are selected for their novelty. Yet, comparing 179 classifiers on 121 datasets shows no statistically significant differences between the top methods [82]. In order to sustain novelty, researchers may be introducing unnecessary complexity into the methods, that do not improve their prediction but rather contribute to technical debt, making systems harder to maintain and deploy[83].

Another metric emphasized is obtaining "state-of-the-art" results, which leads to several of the evaluation problems outlined in Section Evaluations that miss the target. The pressure to publish "good" results can aggravate methodological loopholes[84], for

instance gaming the evaluation in machine learning[85]. It is then all too appealing to find after-the-fact theoretical justifications of positive yet fragile empirical findings. This phenomenon, known as *HARKing* (hypothesizing after the results are known)[86], has been documented in machine learning[87] and computer science in general[62].

Finally, the selection of publications creates the so-called "file drawer problem"[88]: positive results, some due to experimental flukes, are more likely to be published than corresponding negative findings. For example, in 410 most downloaded papers from the ACM, 97% of the papers which used significance testing had a finding with p-value of less than 0.05[62]. It seems highly unlikely that only 3% of the initial working hypotheses—even for impactful work—turned out not confirmed.

Let us improve our publication norms

Fortunately there are various alleys to improve reporting and transparency. For instance, the growing set of open datasets could be leveraged for collaborative work beyond the capacities of a single team[89]. The set of metrics studied could then be broadened, shifting the publication focus away from a single-dimension benchmark. More metrics can indeed help understanding a method's strengths and weaknesses[41,90,91], exploring for instance calibration metrics[28,47,92] or learning curves[93]. The medical-research literature has several reporting guidelines for prediction studies[67,94,95]. They underline many points raised in previous sections: reporting on how representative the study sample is, on the separation between train and test data, on the motivation for the choice of outcome, evaluation metrics, and so forth. Unfortunately, algorithmic research in medical imaging seldom refers to these guidelines.

Methods should be studied on more than prediction performance: reproducibility[81], carbon footprint[96], or a broad evaluation of costs should be put in perspective with the real-world patient outcomes, from a putative clinical use of the algorithms[97].

Preregistration or registered reports can bring more robustness and trust: the motivation and experimental setup of a paper are to be reviewed before empirical results are available, and thus the paper is be accepted before the experiments are run[98]. Translating this idea to machine learning faces the challenge that new data is seldom acquired in a machine learning study, yet it would bring sizeable benefits[62,99].

More generally, accelerating the progress in science calls for accepting that some published findings are sometimes wrong[100]. Popularizing different types of publications may help, for example publishing negative results[101], replication studies[102], commentaries[103] and reflections on the field[68] or the recent NeurIPS Retrospectives workshops. Such initiatives should ideally be led by more established academics, and be welcoming of newcomers[104].

CONCLUSIONS

Despite great promises, the extensive research in medical applications of machine learning seldom achieves a clinical impact. Studying the academic literature and data-science challenges reveals troubling trends: accuracy on diagnostic tasks progresses slower on research cohorts that are closer to real-life settings; methods research is often guided by dataset availability rather than clinical relevance; many developments of model bring improvements smaller than the evaluation errors. We have surveyed challenges of clinical machine-learning research that can explain these difficulties. The challenges start with the choice of datasets, plague model evaluation, and are amplified by publication incentives. Understanding these mechanisms enables us to suggest specific strategies to improve the various steps of the research cycle, promoting publications best practices[105]. None of these strategies are silver-bullet solutions. They rather require changing procedures, norms, and goals. But implementing them will help fulfilling the promises of machine-learning in healthcare: better health outcomes for patients with less burden on the care system.

DATA AVAILABILITY

For reproducibility, all data used in our analyses are available on https://github.com/GaelVaroquaux/ml_med_imaging_failures.

CODE AVAILABILITY

For reproducibility, all code for our analyses is available on https://github.com/GaelVaroquaux/ml_med_imaging_failures.

Received: 21 June 2021; Accepted: 9 March 2022;
Published online: 12 April 2022

REFERENCES

1. Litjens, G. et al. A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).
2. Cheplygina, V., de Bruijne, M. & Pluim, J. P. W. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med. Image Anal.* **54**, 280–296 (2019).
3. Zhou, S. K. et al. A review of deep learning in medical imaging: Image traits, technology trends, case studies with progress highlights, and future promises. Proceedings of the IEEE1-19 (2020).
4. Liu, X. et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The Lancet Digital Health (2019).
5. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
6. Sendak, M. P. et al. A path for translation of machine learning products into healthcare delivery. *Eur. Med. J. Innov.* **10**, 19–00172 (2020).
7. Schwartz, W. B., Patil, R. S. & Szolovits, P. Artificial intelligence in medicine (1987).
8. Roberts, M. et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat. Mach. Intell.* **3**, 199–217 (2021).
9. Willemink, M. J. et al. Preparing medical imaging data for machine learning. Radiology192224 (2020).
10. Mueller, S. G. et al. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimer's Dement.* **1**, 55–66 (2005).
11. Dallora, A. L., Eivazzadeh, S., Mendes, E., Berglund, J. & Anderberg, P. Machine learning and microsimulation techniques on the prognosis of dementia: A systematic literature review. *PLoS ONE* **12**, e0179804 (2017).
12. Arbabshirani, M. R., Plis, S., Sui, J. & Calhoun, V. D. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage* **145**, 137–165 (2017).
13. Sakai, K. & Yamada, K. Machine learning studies on major brain diseases: 5-year trends of 2014–2018. *Jpn. J. Radiol.* **37**, 34–72 (2019).
14. Wen, J. et al. Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation. *Medical Image Analysis* 101694 (2020).
15. Ansart, M. et al. Predicting the progression of mild cognitive impairment using machine learning: a systematic, quantitative and critical review. *Medical Image Analysis* 101848 (2020).
16. Torralba, A. & Efros, A. A. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR)*, 1521–1528 (2011).
17. Dockès, J., Varoquaux, G. & Poline, J.-B. Preventing dataset shift from breaking machine-learning biomarkers. *GigaScience* **10**, giab055 (2021).
18. Zendel, O., Murschitz, M., Humenberger, M. & Herzner, W. How good is my test data? introducing safety analysis for computer vision. *Int. J. Computer Vis.* **125**, 95–109 (2017).
19. Pooch, E. H., Ballester, P. L. & Barros, R. C. Can we trust deep learning models diagnosis? the impact of domain shift in chest radiograph classification. In *MICCAI workshop on Thoracic Image Analysis* (Springer, 2019).
20. Zech, J. R. et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med.* **15**, e1002683 (2018).

21. Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H. & Ferrante, E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences* (2020).
22. Tasdizen, T., Sajjadi, M., Javanmardi, M. & Ramesh, N. Improving the robustness of convolutional networks to appearance variability in biomedical images. In International Symposium on Biomedical Imaging (ISBI), 549–553 (IEEE, 2018).
23. Wachinger, C., Rieckmann, A., Pölsterl, S. & Initiative, A. D. N. et al. Detect and correct bias in multi-site neuroimaging datasets. *Med. Image Anal.* **67**, 101879 (2021).
24. Ashraf, A., Khan, S., Bhagwat, N., Chakravarty, M. & Taati, B. Learning to unlearn: building immunity to dataset bias in medical imaging studies. In *NeurIPS workshop on Machine Learning for Health (ML4H)* (2018).
25. Yu, X., Zheng, H., Liu, C., Huang, Y. & Ding, X. Classify epithelium-stroma in histopathological images based on deep transferable network. *J. Microsc.* **271**, 164–173 (2018).
26. Abbasi-Sureshjani, S., Raumanns, R., Michels, B. E., Schouten, G. & Cheplygina, V. Risk of training diagnostic algorithms on data with demographic bias. In *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, 183–192 (Springer, 2020).
27. Suresh, H. & Guttag, J. V. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002* (2019).
28. Park, S. H. & Han, K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* **286**, 800–809 (2018).
29. Oakden-Rayner, L., Dunnmon, J., Carneiro, G. & Ré, C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In ACM Conference on Health, Inference, and Learning, 151–159 (2020).
30. Winkler, J. K. et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatol.* **155**, 1135–1141 (2019).
31. Joskowicz, L., Cohen, D., Caplan, N. & Sosna, J. Inter-observer variability of manual contour delineation of structures in CT. *Eur. Radiol.* **29**, 1391–1399 (2019).
32. Oakden-Rayner, L. Exploring large-scale public medical image datasets. *Academic Radiol.* **27**, 106–112 (2020).
33. Langley, P. The changing science of machine learning. *Mach. Learn.* **82**, 275–279 (2011).
34. Rabanser, S., Günnemann, S. & Lipton, Z. C. Failing loudly: an empirical study of methods for detecting dataset shift. In *Neural Information Processing Systems (NeurIPS)* (2018).
35. Rädsch, T. et al. What your radiologist might be missing: using machine learning to identify mislabeled instances of X-ray images. In *Hawaii International Conference on System Sciences (HICSS)* (2020).
36. Beyer, L., Hénaff, O. J., Kolesnikov, A., Zhai, X. & Oord, A. v. d. Are we done with ImageNet? *arXiv preprint arXiv:2006.07159* (2020).
37. Gebru, T. et al. Datasheets for datasets. In Workshop on Fairness, Accountability, and Transparency in Machine Learning (2018).
38. Mitchell, M. et al. Model cards for model reporting. In *Fairness, Accountability, and Transparency (FAccT)*, 220–229 (ACM, 2019).
39. Ørting, S. N. et al. A survey of crowdsourcing in medical image analysis. *Hum. Comput.* **7**, 1–26 (2020).
40. Poldrack, R. A., Huckins, G. & Varoquaux, G. Establishment of best practices for evidence for prediction: a review. *JAMA Psychiatry* **77**, 534–540 (2020).
41. Pulini, A. A., Kerr, W. T., Loo, S. K. & Lenartowicz, A. Classification accuracy of neuroimaging biomarkers in attention-deficit/hyperactivity disorder: Effects of sample size and circular analysis. *Biol. Psychiatry.: Cogn. Neurosci. Neuroimaging* **4**, 108–120 (2019).
42. Saeb, S., Lonini, L., Jayaraman, A., Mohr, D. C. & Kording, K. P. The need to approximate the use-case in clinical machine learning. *Gigascience* **6**, gix019 (2017).
43. Hosseini, M. et al. I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data. Neuroscience & Biobehavioral Reviews (2020).
44. Simpson, A. L. et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063* (2019).
45. Rohlfing, T. Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. *IEEE Trans. Med. Imaging* **31**, 153–163 (2011).
46. Maier-Hein, L. et al. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat. Commun.* **9**, 5217 (2018).
47. Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L. & Steyerberg, E. W. Calibration: the Achilles heel of predictive analytics. *BMC Med.* **17**, 1–7 (2019).
48. Wagstaff, K. L. Machine learning that matters. In International Conference on Machine Learning (ICML), 529–536 (2012).
49. Shankar, V. et al. Evaluating machine accuracy on imagenet. In International Conference on Machine Learning (ICML) (2020).
50. Bellamy, D., Celi, L. & Beam, A. L. Evaluating progress on machine learning for longitudinal electronic healthcare data. *arXiv preprint arXiv:2010.01149* (2020).
51. Oliver, A., Odena, A., Raffel, C., Cubuk, E. D. & Goodfellow, I. J. Realistic evaluation of semi-supervised learning algorithms. In Neural Information Processing Systems (NeurIPS) (2018).
52. Dacrema, M. F., Cremonesi, P. & Jannach, D. Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In ACM Conference on Recommender Systems, 101–109 (2019).
53. Musgrave, K., Belongie, S. & Lim, S.-N. A metric learning reality check. In European Conference on Computer Vision, 681–699 (Springer, 2020).
54. Pham, H. V. et al. Problems and opportunities in training deep learning software systems: an analysis of variance. In IEEE/ACM International Conference on Automated Software Engineering, 771–783 (2020).
55. Bouthillier, X. et al. Accounting for variance in machine learning benchmarks. In Machine Learning and Systems (2021).
56. Varoquaux, G. Cross-validation failure: small sample sizes lead to large error bars. *NeuroImage* **180**, 68–77 (2018).
57. Szucs, D. & Ioannidis, J. P. Sample size evolution in neuroimaging research: an evaluation of highly-cited studies (1990–2012) and of latest practices (2017–2018) in high-impact journals. NeuroImage117164 (2020).
58. Roelofs, R. et al. A meta-analysis of overfitting in machine learning. In Neural Information Processing Systems (NeurIPS), 9179–9189 (2019).
59. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **7**, 1–30 (2006).
60. Thompson, W. H., Wright, J., Bissett, P. G. & Poldrack, R. A. Meta-research: dataset decay and the problem of sequential analyses on open datasets. *eLife* **9**, e53498 (2020).
61. Maier-Hein, L. et al. Is the winner really the best? a critical analysis of common research practice in biomedical image analysis competitions. *Nature Communications* (2018).
62. Cockburn, A., Dragicevic, P., Besançon, L. & Gutwin, C. Threats of a replication crisis in empirical computer science. *Commun. ACM* **63**, 70–79 (2020).
63. Gigerenzer, G. Statistical rituals: the replication delusion and how we got there. *Adv. Methods Pract. Psychol. Sci.* **1**, 198–218 (2018).
64. Benavoli, A., Corani, G. & Mangili, F. Should we really use post-hoc tests based on mean-ranks? *J. Mach. Learn. Res.* **17**, 152–161 (2016).
65. Berrar, D. Confidence curves: an alternative to null hypothesis significance testing for the comparison of classifiers. *Mach. Learn.* **106**, 911–949 (2017).
66. Bouthillier, X., Laurent, C. & Vincent, P. Unreproducible research is reproducible. In International Conference on Machine Learning (ICML), 725–734 (2019).
67. Norgeot, B. et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat. Med.* **26**, 1320–1324 (2020).
68. Drummond, C. Machine learning as an experimental science (revisited). In AAAI workshop on evaluation methods for machine learning, 1–5 (2006).
69. Steyerberg, E. W. & Harrell, F. E. Prediction models need appropriate internal, internal–external, and external validation. *J. Clin. Epidemiol.* **69**, 245–247 (2016).
70. Woo, C.-W., Chang, L. J., Lindquist, M. A. & Wager, T. D. Building better biomarkers: brain models in translational neuroimaging. *Nat. Neurosci.* **20**, 365 (2017).
71. Van Calster, B. et al. Reporting and interpreting decision curve analysis: a guide for investigators. *Eur. Urol.* **74**, 796 (2018).
72. Thomas, R. & Uminsky, D. The problem with metrics is a fundamental problem for AI. *arXiv preprint arXiv:2002.08512* (2020).
73. for the Evaluation of Medicinal Products, E. A. Points to consider on switching between superiority and non-inferiority. *Br. J. Clin. Pharmacol.* **52**, 223–228 (2001).
74. D'Agostino Sr, R. B., Massaro, J. M. & Sullivan, L. M. Non-inferiority trials: design concepts and issues–the encounters of academic consultants in statistics. *Stat. Med.* **22**, 169–186 (2003).
75. Christensen, E. Methodology of superiority vs. equivalence trials and non-inferiority trials. *J. Hepatol.* **46**, 947–954 (2007).
76. Hendriksen, J. M., Geersing, G.-J., Moons, K. G. & de Groot, J. A. Diagnostic and prognostic prediction models. *J. Thrombosis Haemost.* **11**, 129–141 (2013).
77. Campbell, M. K., Elbourne, D. R. & Altman, D. G. Consort statement: extension to cluster randomised trials. *BMJ* **328**, 702–708 (2004).
78. Blasini, M., Peiris, N., Wright, T. & Colloca, L. The role of patient–practitioner relationships in placebo and nocebo phenomena. *Int. Rev. Neurobiol.* **139**, 211–231 (2018).
79. Lipton, Z. C. & Steinhardt, J. Troubling trends in machine learning scholarship: some ML papers suffer from flaws that could mislead the public and stymie future research. *Queue* **17**, 45–77 (2019).

80. Tatman, R., VanderPlas, J. & Dane, S. A practical taxonomy of reproducibility for machine learning research. In *ICML workshop on Reproducibility in Machine Learning* (2018).
81. Gundersen, O. E. & Kjensmo, S. State of the art: Reproducibility in artificial intelligence. In *AAAI Conference on Artificial Intelligence* (2018).
82. Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D. & Amorim Fernández-Delgado, D. Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.* **15**, 3133–3181 (2014).
83. Sculley, D. et al. Hidden technical debt in machine learning systems. In Neural Information Processing Systems (NeurIPS), 2503–2511 (2015).
84. Ioannidis, J. P. A. Why most published research findings are false. *PLoS Med.* **2**, e124 (2005).
85. Teney, D. et al. On the value of out-of-distribution testing: an example of Goodhart's Law. In *Neural Information Processing Systems (NeurIPS)* (2020).
86. Kerr, N. L. HARKing: hypothesizing after the results are known. *Personal. Soc. Psychol. Rev.* **2**, 196–217 (1998).
87. Gencoglu, O. et al. HARK side of deep learning–from grad student descent to automated machine learning. *arXiv preprint arXiv:1904.07633* (2019).
88. Rosenthal, R. The file drawer problem and tolerance for null results. *Psychological Bull.* **86**, 638 (1979).
89. Kellmeyer, P. Ethical and legal implications of the methodological crisis in neuroimaging. *Camb. Q. Healthc. Ethics* **26**, 530–554 (2017).
90. Japkowicz, N. & Shah, M. Performance evaluation in machine learning. In *Machine Learning in Radiation Oncology*, 41–56 (Springer, 2015).
91. Santafe, G., Inza, I. & Lozano, J. A. Dealing with the evaluation of supervised classification algorithms. *Artif. Intell. Rev.* **44**, 467–508 (2015).
92. Han, K., Song, K. & Choi, B. W. How to develop, validate, and compare clinical prediction models involving radiological parameters: study design and statistical methods. *Korean J. Radiol.* **17**, 339–350 (2016).
93. Richter, A. N. & Khoshgoftaar, T. M. Sample size determination for biomedical big data with limited labels. *Netw. Modeling Anal. Health Inform. Bioinforma.* **9**, 12 (2020).
94. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): the tripod statement. *J. Br. Surg.* **102**, 148–158 (2015).
95. Wolff, R. F. et al. Probast: a tool to assess the risk of bias and applicability of prediction model studies. *Ann. Intern. Med.* **170**, 51–58 (2019).
96. Henderson, P. et al. Towards the systematic reporting of the energy and carbon footprints of machine learning. *J. Mach. Learn. Res.* **21**, 1–43 (2020).
97. Bowen, A. & Casadevall, A. Increasing disparities between resource inputs and outcomes, as measured by certain health deliverables, in biomedical research. *Proc. Natl Acad. Sci.* **112**, 11335–11340 (2015).
98. Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P. & Willmes, K. Registered reports: realigning incentives in scientific publishing. *Cortex* **66**, A1–A2 (2015).
99. Forde, J. Z. & Paganini, M. The scientific method in the science of machine learning. In ICLR workshop on Debugging Machine Learning Models (2019).
100. Firestein, S.Failure: Why science is so successful (Oxford University Press, 2015).
101. Borji, A. Negative results in computer vision: a perspective. *Image Vis. Comput.* **69**, 1–8 (2018).
102. Voets, M., Møllersen, K. & Bongo, L. A. Replication study: Development and validation of deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *arXiv preprint arXiv:1803.04337* (2018).
103. Wilkinson, J. et al. Time to reality check the promises of machine learning-powered precision medicine. The Lancet Digital Health (2020).
104. Whitaker, K. & Guest, O. #bropenscience is broken science. *Psychologist* **33**, 34–37 (2020).
105. Kakarmath, S. et al. Best practices for authors of healthcare-related artificial intelligence manuscripts. *NPJ Digital Med.* **3**, 134–134 (2020).

ACKNOWLEDGEMENTS

We would like to thank Alexandra Elbakyan for help with the literature review. We thank Pierre Dragicevic for providing feedback on early versions of this manuscript, and Pierre Bartet for comments on the preprint. We also thank the reviewers, Jack Wilkinson and Odd Erik Gundersen, for excellent comments which improved our manuscript. GV acknowledges funding from grant ANR-17-CE23-0018, DirtyData.

AUTHOR CONTRIBUTIONS

Both V.C. and G.V. collected the data; conceived, designed, and performed the analysis; reviewed the literature; and wrote the paper.

COMPETING INTERESTS

The authors declare that there are no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-022-00592-y.

Correspondence and requests for materials should be addressed to Gaël. Varoquaux or Veronika Cheplygina.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.