# How I failed machine learning in medical imaging - shortcomings and recommendations

Gaël Varoquaux*, Veronika Cheplygina†
*INRIA, France
†IT University of Copenhagen, Denmark

**Disclaimer: this is a working paper, and represents research in progress. For missing references and other comments or questions, please email us at gael.varoquaux@inria.fr and vech@itu.dk**

***Abstract*—Medical imaging is an important research field with many opportunities for improving patients' health. However, there are a number of challenges that are slowing down the progress of the field as a whole, such optimizing for publication. In this paper we reviewed several problems related to choosing datasets, methods, evaluation metrics, and publication strategies. With a review of literature and our own analysis, we show that at every step, potential biases can creep in. On a positive note, we also see that initiatives to counteract these problems are already being started. Finally we provide a broad range of recommendations on how to further these address problems in the future. For reproducibility, data and code for our analyses are available on https://github.com/GaelVaroquaux/ml_med_imaging_failures.**

## I. Introduction

The great process in machine learning opens the door to many improvements in medical image processing [Litjens et al., 2017, Cheplygina et al., 2019, Zhou et al., 2020]. For example, to diagnose various conditions from medical images, ML algorithms have been shown to perform on par with medical experts [see Liu et al., 2019, for a recent overview]. Software applications are starting to be certified for clinical use [Topol, 2019, Sendak et al., 2020].

The stakes are high, and there is a staggering amount of research on machine learning for medical images, as many recent surveys show. This growth does not inherently lead to clinical progress. The higher volume of research can be aligned with the academic incentives rather than the needs of clinicians and patients. As an 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.

In this paper, we explore avenues to improve clinical impact of machine learning research in medical imaging. After sketching the situation, documenting uneven progress, we study a number of failures we see in some medical imaging papers, which occur at different steps of the "publishing lifecycle":

- What data to use (Section III)
- What method to use and how to evaluate them (Section IV)
- How to publish the results (Section V)

In each section we first discuss the problems, supported with evidence from previous research as well as our own analyses of recent medical imaging work. We then discuss a number of steps to improve the situation, sometimes borrowed from related communities. We hope that these ideas will help shape a research community even more effective at addressing real-world medical-imaging problems.

## II. It's not all about larger datasets

The availability of large labeled datasets has enabled solving difficult artificial intelligence problems, such as natural scene understanding in computer vision [Russakovsky et al., 2015]. As a result, there is widespread hope that similar progress will happen in medical applications: with large datasets, algorithm research will eventually solve a clinical problem posed as discrimination task. 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 often failed to 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. Hence, efforts have been dedicated to acquire large brain-imaging cohorts of aging individuals at risk of developing AD, on which early biomarkers can be developed using machine learning [Mueller et al., 2005]. 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, as visible in Figure 1a, built with a meta-analysis compiling 478 studies from 6 systematic reviews [Dallora et al., 2017, Arbabshirani et al., 2017, Liu et al., 2019, Sakai and Yamada, 2019, Wen et al., 2020, Ansart et al., 2020].

However, the increase in data size 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 (Figure 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. However, research efforts across time lead to improvements even on large, heterogeneous cohorts (Figure 1c), as studies published later show improvements for large sample sizes.

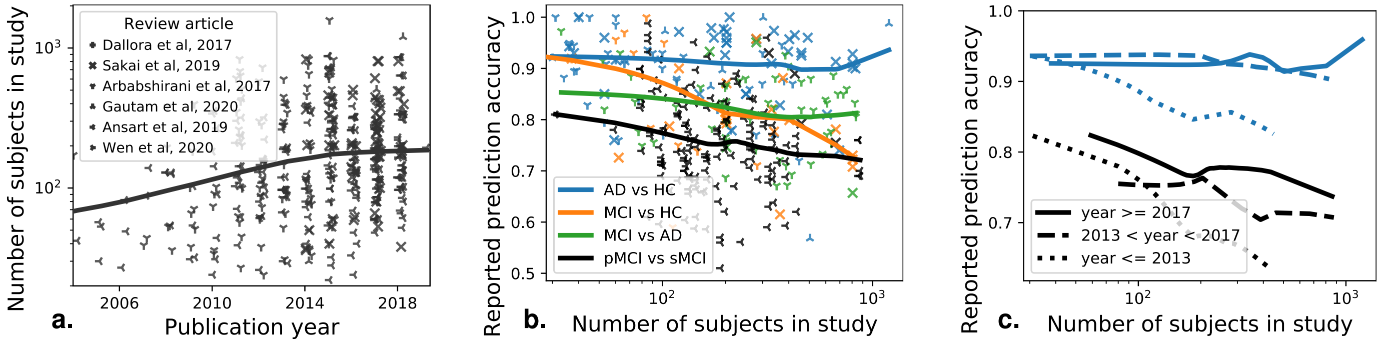

Fig. 1. **Bigger brain-imaging datasets do not suffice for better machine-learning based diagnosis of Alzheimer's**. A meta-analysis across 6 review papers, covering more than 500 individual publications. The machine-learning problem is typically formulated as distinguishing various related clinical 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 distingishing studies published in different years.

## III. Data, an imperfect window on the clinic

### A. *Datasets reflect an application only partly*

Available datasets only partially reflect the clinical situation for a particular medical condition, leading to dataset bias. This is an important problem, especially given that the investigator may be unaware of this dataset bias. Dataset bias occurs when the data used to build the decision model ( the training data), has a different distribution than the data representing the population on which it should be applied (the test data). This happens for example if the training data was acquired with a different type of scanner. Dataset bias was initially studied in computer vision [Torralba and Efros, 2011, Recht et al., 2019, 2018, among others]: images of cars from a given benchmark dataset are actually not representative of more general images of cars. As a result, algorithms which score high in benchmarks can perform poorly in real world scenarios [Zendel et al., 2017]. In medical imaging, dataset bias has been demonstrated in chest X-rays [Pooch et al., 2019, Zech et al., 2018, Larrazabal et al., 2020], retinal imaging [Tasdizen et al., 2018], brain imaging [Wachinger et al., 2018, Ashraf et al., 2018], histopathology [Yu et al., 2018], or dermatology [Abbasi-Sureshjani et al., 2020].

There are many potential sources dataset bias in medical imaging, introduced at different phases of the modeling process [Suresh and Guttag, 2019]. First, a cohort may not appropriately represent the range of possible patients and disease manifestations for the image analysis task at hand, a bias sometimes called *spectrum bias* [Park and Han, 2018]. Such bias are typically revealed by training and testing a model across datasets from different sources, and observing a performance drop across sources. A detrimental consequence is that model performance can be overestimated for different groups, for example between male and female subjects [Abbasi-Sureshjani et al., 2020, Larrazabal et al., 2020]. Yet medical imaging publications seldom report the demographics of the data [Abbasi-Sureshjani et al., 2020].

Imaging devices or procedures may lead to specific measurement biases. A bias particularly harmful to clinically-relevant automated diagnosis is when the data captures medical interventions. For instance, Oakden-Rayner et al. [2020] showed that on chest X-ray images, the "pneumothorax" condition sometimes show a chest drain, which is a treatment for this condition, and which would not yet be present before diagnosis. Similar spurious correlations can appear in skin lesion images due to markings placed by dermatologists next to the lesions [Winkler et al., 2019].

Labeling errors can also introduce biases. Expert human annotators may give different labels with systematic biases [Joskowicz et al., 2019], but multiple annotators are seldom available. Using automatic methods to extract labels from patient reports can also lead to problems, as Oakden-Rayner [2020] shows for the chest X-ray data. For example, findings which have been known from previous reports on the patient, may not be stated during a follow-up scan, and therefore seen as "negative" for that category by the automatic labeling procedure.

### B. *Dataset availability distorts research*

The availability of public datasets can influence which applications are studied more frequently. To evaluate this, we compare two different applications within medical imag-

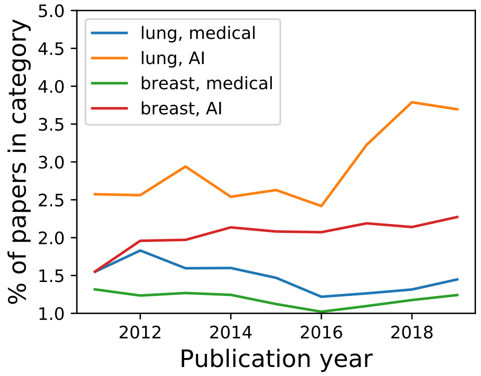

Fig. 2. Differences between popularity of applications. We show the percentage of papers on lung vs breast cancer, for medical oncology and artificial intelligence. The percentages are relatively constant, except lung cancer in AI, which shows an increase after 2016.

ing: detecting lung nodules, and detecting breast tumors in radiological images. We choose these applications due to the wide availability of various lung datasets on Kaggle or grand-challenge.org, contrasted with (to our knowledge) only one challenge focusing on mammograms.

We then quantify the prevalence of these topics in general medical literature, and in the field of machine learning. We use the Dimensions.AI app [Mori and Taylor, 2018], querying the titles and abstracts of papers, with the following two queries:

- lung AND (tumor OR nodule) AND (scan OR image)
- breast AND (tumor OR nodule) AND (scan OR image)

We do this for two categories, which are the largest subcategories within top-level categories "medical sciences" and "information computing":

- 1112 Oncology and Carcinogenesis
- 0801 Artificial Intelligence and Image Processing

We then normalize the number of papers per year, by the total number of papers for the "cancer AND (scan OR image)" query in the respective categories (1112 Oncology or 0801 AI). The graph in Figure 2 shows that the percentages are relatively constant, except lung imaging in the AI category, which shows a substantial increase in 2016. We suspect that the Kaggle lung challenges published around that time contributed to this disproportional increase.

### C. Let us build awareness of data limitations

We need to critically think about our choice of datasets if we want to address these problems. This can mean evaluating our dataset choice on a project level, i.e. which datasets to select for a paper or a challenge, or on a higher level, i.e. which datasets we work on as a community.

On a 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 [Willemink et al., 2020] but this may not always be possible. However, existing datasets can still be critically evaluated for the presence of dataset shift [Rabanser et al., 2018], hidden subgroups (not reflected in the meta-data) [Oakden-Rayner et al., 2020], mislabeled instances [Rädsch et al., 2020] or other biases [Suresh and Guttag, 2019]. A checklist for structurally evaluating computer vision datasets for such problems is presented in [Zendel et al., 2017]. When problems are discovered, it might be worth investing time into relabeling a subset of the data [Beyer et al., 2020].

On a higher level, we should strive to be aware of the limitations of datasets as a community. When creating a new dataset or challenge, it is advisable to document the dataset with its characteristics, and thus possible model limitations. Possibilities include data sheets [Gebru et al., 2018], which describe the data collection procedure, and model cards [Mitchell et al., 2019], which describe the choices made to train a model (including the data).

Meta-analyses which look at how dataset use in different areas evolves, are another way to reflect on what we are focusing on as a field overall. For example, a survey of crowdsourcing in medical imaging [Ørting et al., 2020] shows a different distribution of applications than surveys focusing on machine learning methods [Litjens et al., 2017, Cheplygina et al., 2019], which could point at under-addressed problems. Similarly, we could compare more clinically oriented venues to more technical venues, and so forth.

## IV. Evaluations that miss the target

### A. Metrics that do not reflect what we want

Evaluating models requires choosing a suitable metric, however, our understanding of "suitable" may change over time. As a classical example, the image similarity metric which was widely used to evaluate the quality of image registration algorithms, was later shown to be ineffective as high values could be obtained for scrambled images [Rohlfing, 2011].

In medical image segmentation, Maier-Hein et al. [2018a] review 150 publicly available biomedical image analysis challenges conducted up to 2016 and show that the typical metrics used to rank algorithms are sensitive different variants of the same metric, which may not be suitable if we want to create a robust ranking of algorithms.

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 [Han et al., 2016, Park and Han, 2018].

Finally, the metrics we use may not be synonymous with practical improvement [Wagstaff, 2012, Shankar et al., 2020]. For example, Shankar et al. [2020] shows that typical metrics in computer vision do not reflect what we might find important about image recognition, such as robustness to out-of-distribution examples. Similarly, in a medical imaging application, improvements in traditional metrics may not necessarily translate to different clinical outcomes.

### B. Improper evaluation procedures

For the evaluation procedure, it is generally accepted that that algorithms need to be trained and tested on different sets of data. However, overfitting may still occur, leading to overoptimistic results. For example Pulini et al. [2019] show that studies that classify ADHD based on neuroimaging can engage in circular analysis, where feature selection is done on the full dataset, before cross-validation.

A related but more difficult to detect issue, is what we call "overfitting by observer". Even when cross-validation is carried out for all steps of the method, overfitting may still occur by the researcher observing the cross-validation performance, and adjusting the method. An excellent simulation of this phenomenon can be found in [Hosseini et al., 2020].

### C. Incorrectly chosen baselines

To claim superiority, novel algorithms may be compared to baselines, but these baselines may be poorly chosen. There are two ways this can go wrong: choosing baselines that are not as strong as they could be (under-powered) baselines, and not choosing simple but effective baselines.

When under-powered baselines are chosen, the algorithm results may seem more favorable, creating an illusion of progress. This is illustrated in healthcare data [Bellamy et al., 2020], semi-supervised learning [Oliver et al., 2018], recommendation systems [Dacrema et al., 2019], metric learning [Musgrave et al., 2020] and deep learning more generally [Marcus, 2018] (and references therein).

The opposite problem is not including simple baselines, which may be effective for the problem at hand. For example, Brendel and Bethge [2019] show that some convolutional neural networks (CNNs) can be approximated by a bag-of-feature approach, and Wen et al. [2020] show that CNNs do not outperform support vector machines.

### *D. Statistical significance not tested, or misunderstood*

Machine learning experiments drive the progress of the field via empirical results. However, this evidence is by nature noisy: results may depend on which specific samples were used to train the models, the random seeds used to initialize them, small differences in hyper-parameters [Bouthillier et al., 2019, 2021]. Separating out noise from findings requires statistical testing, however, there is a lack of well-established good practices for predictive modeling.

Problems arising from drawing conclusions on too small sample sizes are well documented [Varoquaux, 2018, Airola et al., 2009]. Indeed, predictive modeling studies require many samples, more than conventional inferential studies, else the measured prediction accuracy may be a distant estimation of real-life performance. However, sample sizes are evolving slowly [Szucs and Ioannidis, 2020], and few studies justify their sample size [Larson and Carbine, 2017]. On a positive note, Roelofs et al. [2019] present a meta-analysis of public vs private leader boards on Kaggle for different applications, but conclude show that with "large enough" test data, overfitting is less of an issue - the two test sets that show substantial overfitting have either 90 or 4000 (but derived from 7 human participants) examples.

Another challenge is that strong validation of a method requires it to be robust to details of the training data. Hence it should go beyond evidence on a single dataset, and rather strive for statistically significant consensus across multiple datasets [Demšar, 2006]. The existing 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 [Thompson et al., 2020], in an implicit multiple comparison problem.

And yet, medical imaging research seldom analyzes how likely empirical results are to be due to chance: only 6% of segmentation challenges surveyed [Maier-Hein et al., 2018b], and 15% out of 410 computer science papers published by ACM [Cockburn et al., 2020] used a statistical test.

However, null-hypothesis testing does not bring a full answer, as outlined in [Demšar, 2008] by the very author of [Demšar, 2006]. Null-hypothesis tests are often misinterpreted [Gigerenzer, 2018], often along two specific 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 [Benavoli et al., 2016, Berrar, 2017].

A last challenge is that publication incentives erode statistical control, as discussed in subsection V-B.

### *E. Evaluation error is often larger than algorithmic improvements*

As a result of the focus on outperforming performance metrics, we may be facing a situation of diminishing returns, where relatively more effort is needed to achieve relatively smaller performance gains. Furthermore, these gains may not be practically significant, but part of the background noise. To evaluate whether this situation is happening, we need to model the distribution of typical performance improvements. A novel algorithm's performance gain can then be compared to this background distribution.

To get data on performance improvements by different algorithms, we looked at four Kaggle competitions, two focused on classification and two focused on segmentation. The details are given in Table I.

For each competition, we looked at the public and private leaderboards, extracting the following information:

- Differences $d_i$, defined by the difference of the $i$-th algorithm between the public and private leaderboard
- Distribution of $d_i$'s per competition, its mean and standard deviation
- The interval $t_{10}$, defined by the difference between the best algorithm, and the "top 10%" algorithm

Our findings are shown in Figure 3. For brain MRI and lung cancer diagnosis, we see a relatively wide spread distributions of performance differences. These distributions are centered at zero, which means there is no overfitting on the public datasets. The interval $t_{10}$ is well within this distribution, which means that this difference could also be explained by noise. For the lung tumor segmentation challenge, the distribution of differences is narrow, but offset to the left, suggesting overfitting on the public set. Again $t_{10}$ is small, suggesting the diminishing returns scenario.

On the other hand, in the nerve segmentation challenge we see a moderately spread distribution that is slightly offset to the right, indicating better performance on the private dataset. In contrast to the other three challenges, the difference between the best and the "top 10% algorithm" is much larger than what could be explained by noise, and thus truly meaningful.

### *F. Let us redefine evaluation*

*a) Higher standards for benchmarking:* Good machine-learning benchmarks are more difficult than they may seem. Various recent publications outline best practices for medical machine learning evaluation [Park and Han, 2018, England and Cheng, 2019, Poldrack et al., 2020, Norgeot et al., 2020], which we summarize below:

- Safeguarding from data leakage by separating out all test data from the start, before any data transformation.

| Description | URL | Incentive | Test size | Entries |
|---|---|---|---|---|
| Schizophrenia classification in MR scans | https://www.kaggle.com/c/mlsp-2014-mri/overview | Publications | 120K | 313 |
| Pneumothorax segmentation in X-rays | https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation | 30K USD | max 6K | 350 |
| Nerve segmentation in ultrasound images | https://www.kaggle.com/c/ultrasound-nerve-segmentation | 100K USD | 5.5K | 922 |
| Lung cancer detection in CT scans | https://www.kaggle.com/c/data-science-bowl-2017 | 1M USD | Not available | 394 |

TABLE I

DETAILS OF KAGGLE CHALLENGES USED FOR OUR ANALYSIS. THE TEST SIZE SHOWS THE NUMBER OF TEST IMAGES PROVIDED, AND THE NUMBER OF ENTRIES CORRESPONDS TO THE NUMBER OF RESULTS ON THE PRIVATE LEADERBOARD.

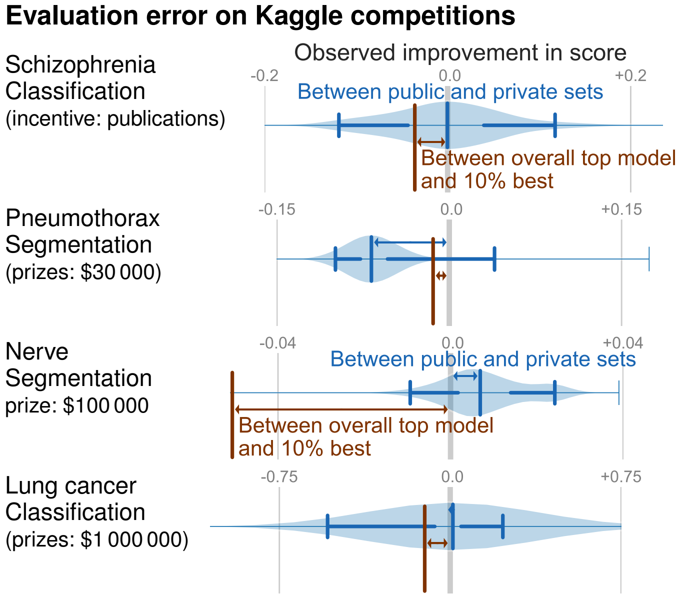

Fig. 3. **Kaggle competitions: shifts from public to private set compared to improvement across top-most 10% models** on 4 medical-imaging competitions with significant incentives. The blue violin plot gives the distribution of differences between public and private leaderboards (positive means that private leaderboard is better than public leaderboard). A systematic shift between public and private set indicates an overfit or a dataset shift. The width of this distribution gives the intrinsic evaluation noise of the competition. The brown bar is $t_{10}$, 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 public and private set: if it is smaller, the 10% best models reached diminishing returns and did not lead to a actual improvement on new data.

- A documented way of selecting model hyper-parameters (including architectural parameters for neural networks), 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, preferably with a motivated sample size determination [Willemink et al., 2020, Richter and Khoshgoftaar, 2020].
- 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 [Steyerberg and Harrell, 2016, Woo et al., 2017].
- Strong baselines that reflect the state of the art of machine-learning research on the given topic, but also historical solutions including clinical methodologies not necessarily relying on medical imaging.
- A discussion the variability of the results with regards to arbitrary choices (random seeds) and data sources with an eye on statistical significance.
- Using different quantitative metrics to capture the different aspects of the clinical problem and relating them to relevant clinical performance metric.
- Adding qualitative accounts and involving groups that will be most affected by the application in the metric design [Thomas and Uminsky, 2020].

*b) More than beating the benchmark:* As already suggested above, we want to look at more than just outperforming the benchmark, even if this is done with proper validation and statistical significance testing. One point of view is that rejecting a null is not sufficient, and that 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 *non-inferiority tests* used in clinical trials [D'Agostino Sr et al., 2003, Christensen, 2007, Schumi and Wittes, 2011, Bokai et al., 2017]. For predictive modeling benchmarks, it amounts to comparing the observed improvement to variation of the results due to arbitrary choices such as data sampling or random seeds [Bouthillier et al., 2021].

Challenges offer a way of limiting the winner's curse, but ideally we want challenges not to only focus on the winner. Instead, much more knowledge could be extracted by comparing the competing methods and analysing the determinants of success, as well as analyzing failure cases. The MICCAI working group on challenges is already making steps towards this.

## V. PUBLISHING, DISTORTED INCENTIVES

### A. No incentive for clarity

The publication process does not create incentives for clarity. Lipton and Steinhardt [2019] describe various problems where this becomes apparent, such as unnecessary "mathiness" of papers and suggestive language (such as "human-level performance"). There are even a number of memes about this, such as [Von Bearnensquash, 2010] show that including equations increases the chance a paper is accepted to the main conference (rather than a less prestigious workshop).

Additionally, sometimes important details may be not reported, from ablation experiments which show what part of the method led to the most improvements [Lipton and Steinhardt, 2019], to reporting how algorithms were evaluated in a challenge [Maier-Hein et al., 2018a]. This in turn influences the reproducibility of the results, which can be viewed on several levels: from reproducing the exact results to being able to draw the same conclusions [McDermott et al., 2019, Tatman et al., 2018, Gundersen and Kjensmo, 2018].

### *B. Optimizing for publication*

As researchers we want to say our main goal is to solve a particular scientific problem, however, the reality of the culture we exist in, can distort this objective. A good summary of this problem is Goodhart's Law - *when a measure becomes a target, it ceases to be a good measure*. Goodhart's Law manifests itself in several steps of the publishing life-cycle.

One metric that is emphasized is the publication of novel methods. Previous research comparing 179 classifiers on 121 datasets has shown that while some classifiers are generally more accurate, there are no statistically significant differences between the top methods [Fernández-Delgado et al., 2014]. In order to sustain novelty, researchers may be introducing unnecessary complexity into the method, contributing to technical debt˜[Sculley et al., 2015] - creating unnecessary complex, and possibly less stable or understandable code.

Another metric that is emphasized is obtaining "state-of-the-art" results, which leads to several of the evaluation problems outlined in Section IV. This can lead to a *HARKing* (hypothesizing after the results are known) [Kerr, 1998]. This phenomenon is also documented in machine learning [Gencoglu et al., 2019] and computer science in general [Cockburn et al., 2020]. The pressure to publish "good" results can also incentivize questionable data analysis methods [Ioannidis, 2005] or intentionally gaming the evaluation metrics [Teney et al., 2020].

At publication time, we are also confronted with the file drawer problem Rosenthal [1979], where "positive" or "significant" results are more likely to be published. For example, Cockburn et al. [2020] finds that 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, where in reality a uniform distribution of p-values would be expected.

### *C. Let us improve our publication norms*

Fortunately there are various ways in which the reporting and transparency can be improved. Although open datasets and challenges are more common now, collaboration within such initiatives could still be improved. This would for example allow analyses that single teams are not able to do [Kellmeyer, 2017].

Expanding the metrics we examine would be another way to shift the focus of publications. There are more metrics that may be important for understanding a method's strengths and weaknesses, such as calibration metrics [Park and Han, 2018, Han et al., 2016] and learning curves [Richter and Khoshgoftaar, 2020, Beleites et al., 2013]. Good tutorials on metrics and their estimation can be found in [Japkowicz and Shah, 2015, Santafe et al., 2015, Pulini et al., 2019].

Beyond accuracy, we might want to think about how the method affects the outside world, such as reproducibility [Gundersen and Kjensmo, 2018] or the carbon footprint of training the models [Henderson et al., 2020, Anthony et al., 2020]. In a similar vein to [Bowen and Casadevall, 2015], we could also consider comparing the costs (such as funding and resources) to the real-world patient outcomes, such as algorithms used in the clinic.

We could even consider publishing before the results are available, through preregistration or registered reports. The idea is that the motivation and experimental setup of a paper will be reviewed, and thus a paper will be accepted before the data is collected. This may not be one-to-one translatable to machine learning, although some discussions about this have been started [Forde and Paganini, 2019, Cockburn et al., 2020].

More generally, it should not be a bad thing to realize that scientists are sometimes wrong - progress in science even depends on it [Firestein, 2015]. Specific steps to take could be introducing different types of publications or venues. For example publishing negative results [Borji, 2018], replication studies [Voets et al., 2018], commentaries [Wilkinson et al., 2020] and retrospectives (such as the Retrospectives workshop at NeurIPS 2019 and 2020).

## VI. Conclusions

We provide a broad overview of problems which may be slowing down in medical imaging, and related computational fields in general, based on both literature and our own analysis. Our first analysis shows that dataset size is not everything, and while datasets are slowly getting larger, predictive performance is not. Our second analysis shows that the availability of datasets might influence what medical imaging choose to work on, possibly moving attention away from other unsolved problems. Our third analysis shows that outperforming a state-of-the-art method may not always be meaningful, and thus may create an illusion of progress. Next to this, we also provide a broad range of strategies to address this situation, some of which are already being introduced. We hope that these suggestions will be useful to practitioners and medical imaging and related fields.

## Acknowledgements

We would like to thank Alexandra Elbakyan for help with the literature review.

## References

S. Abbasi-Sureshjani, R. Raumanns, B. E. Michels, G. Schouten, and V. Cheplygina. Risk of training diagnostic algorithms on data with demographic bias. In *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, pages 183–192. Springer, 2020.

A. Airola, T. Pahikkala, W. Waegeman, B. De Baets, and T. Salakoski. A comparison of auc estimators in small-sample studies. In *Machine Learning in Systems Biology*, pages 3–13, 2009.

M. Ansart, S. Epelbaum, G. Bassignana, A. Bône, S. Bottani, T. Cattai, R. Couronne, J. Faouzi, I. Koval, M. Louis, et al. Predicting the progression of mild cognitive impairment using machine learning: A systematic, quantitative and critical review. *Medical Image Analysis*, page 101848, 2020.

L. F. W. Anthony, B. Kanding, and R. Selvan. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. In *ICML workshop on Challenges in Deploying and Monitoring Machine Learning Systems*. 2020.

M. R. Arbabshirani, S. Plis, J. Sui, and V. D. Calhoun. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*, 145:137–165, 2017.

A. Ashraf, S. Khan, N. Bhagwat, M. Chakravarty, and B. Taati. Learning to unlearn: Building immunity to dataset bias in medical imaging studies. In *NeurIPS workshop on Machine Learning for Health (ML4H)*. 2018.

C. Beleites, U. Neugebauer, T. Bocklitz, C. Krafft, and J. Popp. Sample size planning for classification models. *Analytica Chimica Acta*, 760:25–33, 2013.

D. Bellamy, L. Celi, and A. L. Beam. Evaluating progress on machine learning for longitudinal electronic healthcare data. *arXiv preprint arXiv:2010.01149*, 2020.

A. Benavoli, G. Corani, and F. Mangili. Should we really use post-hoc tests based on mean-ranks? *The Journal of Machine Learning Research*, 17(1):152–161, 2016.

D. Berrar. Confidence curves: an alternative to null hypothesis significance testing for the comparison of classifiers. *Machine Learning*, 106(6):911–949, 2017.

L. Beyer, O. J. Hénaff, A. Kolesnikov, X. Zhai, and A. v. d. Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

W. Bokai, W. Hongyue, M. Xin, and F. Changyong. Comparisons of superiority, non-inferiority, and equivalence trials. *Shanghai Archives of Psychiatry*, 29(6):385, 2017.

A. Borji. Negative results in computer vision: A perspective. *Image and Vision Computing*, 69:1–8, 2018.

X. Bouthillier, C. Laurent, and P. Vincent. Unreproducible research is reproducible. In *International Conference on Machine Learning (ICML)*, pages 725–734, 2019.

X. Bouthillier, P. Delaunay, M. Bronzi, A. Trofimov, B. Nichyporuk, J. Szeto, N. Mohammadi Sepahvand, E. Raff, K. Madan, V. Voleti, S. E. Kahou, V. Michalski, T. Arbel, C. Pal, G. Varoquaux, and P. Vincent. Accounting for variance in machine learning benchmarks. In *Machine Learning and Systems*, 2021.

A. Bowen and A. Casadevall. Increasing disparities between resource inputs and outcomes, as measured by certain health deliverables, in biomedical research. *Proceedings of the National Academy of Sciences*, 112(36):11335–11340, 2015.

W. Brendel and M. Bethge. Approximating CNNs with bag-of-local-features models works surprisingly well on imagenet. In *International Conference on Learning Representations (ICLR)*. 2019.

V. Cheplygina, M. de Bruijne, and J. P. W. Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, 54:280–296, 2019.

E. Christensen. Methodology of superiority vs. equivalence trials and non-inferiority trials. *Journal of Hepatology*, 46 (5):947–954, 2007.

A. Cockburn, P. Dragicevic, L. Besançon, and C. Gutwin. Threats of a replication crisis in empirical computer science. *Communications of the ACM*, 63(8):70–79, 2020.

M. F. Dacrema, P. Cremonesi, and D. Jannach. Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *ACM Conference on Recommender Systems*, pages 101–109, 2019.

R. B. D'Agostino Sr, J. M. Massaro, and L. M. Sullivan. Non-inferiority trials: design concepts and issues–the encounters of academic consultants in statistics. *Statistics in Medicine*, 22(2):169–186, 2003.

A. L. Dallora, S. Eivazzadeh, E. Mendes, J. Berglund, and P. Anderberg. Machine learning and microsimulation techniques on the prognosis of dementia: A systematic literature review. *PLoS ONE*, 12(6):e0179804, 2017.

J. Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7: 1–30, 2006.

J. Demšar. On the appropriateness of statistical tests in machine learning. In *ICML workshop on Evaluation Methods for Machine Learning*, page 65, 2008.

J. R. England and P. M. Cheng. Artificial intelligence for medical image analysis: a guide for authors and reviewers. *American Journal of Roentgenology*, 212(3):513–519, 2019.

M. Fernández-Delgado, E. Cernadas, S. Barro, D. Amorim, and D. Amorim Fernández-Delgado. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*, 15:3133–3181, 2014. ISSN 1532-4435.

S. Firestein. *Failure: Why science is so successful*. Oxford University Press, 2015.

J. Z. Forde and M. Paganini. The scientific method in the science of machine learning. In *ICLR workshop on Debugging Machine Learning Models*. 2019.

T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. M. Wallach, H. D. III, and K. Crawford. Datasheets for datasets. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning*. 2018.

O. Gencoglu, M. van Gils, E. Guldogan, C. Morikawa, M. Süzen, M. Gruber, J. Leinonen, and H. Huttunen. Hark side of deep learning–from grad student descent to automated machine learning. *arXiv preprint arXiv:1904.07633*, 2019.

G. Gigerenzer. Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science*, 1(2):198–218, 2018.

O. E. Gundersen and S. Kjensmo. State of the art: Reproducibility in artificial intelligence. In *AAAI Conference on Artificial Intelligence*, 2018.

K. Han, K. Song, and B. W. Choi. How to develop, validate, and compare clinical prediction models involving radiological parameters: study design and statistical methods. *Korean Journal of Radiology*, 17(3):339–350, 2016.

P. Henderson, J. Hu, J. Romoff, E. Brunskill, D. Jurafsky, and J. Pineau. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43, 2020.

M. Hosseini, M. Powell, J. Collins, C. Callahan-Flintoft, W. Jones, H. Bowman, and B. Wyble. I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data. *Neuroscience & Biobehavioral Reviews*, 2020.

J. P. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, 2005.

N. Japkowicz and M. Shah. Performance evaluation in machine learning. In *Machine Learning in Radiation Oncology*, pages 41–56. Springer, 2015.

L. Joskowicz, D. Cohen, N. Caplan, and J. Sosna. Inter-observer variability of manual contour delineation of structures in CT. *European Radiology*, 29(3):1391–1399, 2019.

P. Kellmeyer. Ethical and legal implications of the methodological crisis in neuroimaging. *Cambridge Quarterly of Healthcare Ethics*, 26(4):530–554, 2017.

N. L. Kerr. Harking: Hypothesizing after the results are known. *Personality and social psychology review*, 2(3):196–217, 1998.

A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 2020.

M. J. Larson and K. A. Carbine. Sample size calculations in human electrophysiology (eeg and erp) studies: A systematic review and recommendations for increased rigor. *International Journal of Psychophysiology*, 111:33–41, 2017.

Z. C. Lipton and J. Steinhardt. Troubling trends in machine learning scholarship: Some ml papers suffer from flaws that could mislead the public and stymie future research. *Queue*, 17(1):45–77, 2019.

G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. Van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

X. Liu, L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, 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.

L. Maier-Hein, M. Eisenmann, A. Reinke, S. Onogur, M. Stankovic, P. Scholz, T. Arbel, H. Bogunovic, A. P. Bradley, A. Carass, et al. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature Communications*, 9(1):5217, 2018a.

L. Maier-Hein, M. Eisenmann, A. Reinke, S. Onogur, M. Stankovic, P. Scholz, T. Arbel, H. Bogunovic, A. P. Bradley, A. Carass, et al. Is the winner really the best? a critical analysis of common research practice in biomedical image analysis competitions. *Nature Communications*, 2018b.

G. Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.

M. McDermott, S. Wang, N. Marinsek, R. Ranganath, M. Ghassemi, and L. Foschini. Reproducibility in machine learning for health. In *ICLR workshop on Reproducibility in Machine Learning*, 2019.

M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. In *Fairness, Accountability, and Transparency (FAccT)*, pages 220–229. ACM, 2019.

A. Mori and M. Taylor. Dimensions metrics API reference & getting started. *Digital Science & Research solutions*, 2018.

S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. R. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett. Ways toward an early diagnosis in alzheimer's disease: the alzheimer's disease neuroimaging initiative (adni). *Alzheimer's & Dementia*, 1(1):55–66, 2005.

K. Musgrave, S. Belongie, and S.-N. Lim. A metric learning reality check. In *European Conference on Computer Vision*, pages 681–699. Springer, 2020.

B. Norgeot, G. Quer, B. K. Beaulieu-Jones, A. Torkamani, R. Dias, M. Gianfrancesco, R. Arnaout, I. S. Kohane, S. Saria, E. Topol, et al. Minimum information about clinical artificial intelligence modeling: the mi-claim checklist. *Nature Medicine*, 26(9):1320–1324, 2020.

L. Oakden-Rayner. Exploring large-scale public medical image datasets. *Academic Radiology*, 27(1):106–112, 2020.

L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *ACM Conference on Health, Inference, and Learning*, pages 151–159, 2020.

A. Oliver, A. Odena, C. Raffel, E. D. Cubuk, and I. J. Goodfellow. Realistic evaluation of semi-supervised learning algorithms. In *Neural Information Processing Systems (NeurIPS)*, 2018.

S. N. Ørting, A. Doyle, A. van Hilten, M. Hirth, O. Inel, C. R. Madan, P. Mavridis, H. Spiers, and V. Cheplygina. A survey of crowdsourcing in medical image analysis. *Human Computation*, 7:1–26, 2020.

S. H. Park and K. Han. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology*, 286(3):800–809, 2018.

R. A. Poldrack, G. Huckins, and G. Varoquaux. Establishment of best practices for evidence for prediction: a review. *JAMA Psychiatry*, 77(5):534–540, 2020.

E. H. Pooch, P. L. Ballester, and R. C. Barros. 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.

A. A. Pulini, W. T. Kerr, S. K. Loo, and A. Lenartowicz. Classification accuracy of neuroimaging biomarkers in attention-deficit/hyperactivity disorder: Effects of sample size and circular analysis. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 4(2):108–120, 2019.

S. Rabanser, S. Günnemann, and Z. C. Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. In *Neural Information Processing Systems (NeurIPS)*. 2018.

T. Rädsch, S. Eckhardt, F. Leiser, K. D. Pandl, S. Thiebes, and A. Sunyaev. 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.

B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do CIFAR-10 classifiers generalize to CIFAR-10? *arXiv preprint* *arXiv:1806.00451*, 2018.

B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning (ICML)*, pages 5389–5400, 2019.

A. N. Richter and T. M. Khoshgoftaar. Sample size determination for biomedical big data with limited labels. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 9(1):12, 2020.

R. Roelofs, V. Shankar, B. Recht, S. Fridovich-Keil, M. Hardt, J. Miller, and L. Schmidt. A meta-analysis of overfitting in machine learning. In *Neural Information Processing Systems (NeurIPS)*, pages 9179–9189, 2019.

T. Rohlfing. Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. *IEEE Transactions on Medical Imaging*, 31(2):153–163, 2011.

R. Rosenthal. The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3):638, 1979.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

K. Sakai and K. Yamada. Machine learning studies on major brain diseases: 5-year trends of 2014–2018. *Japanese Journal of Radiology*, 37(1):34–72, 2019.

G. Santafe, I. Inza, and J. A. Lozano. Dealing with the evaluation of supervised classification algorithms. *Artificial Intelligence Review*, 44(4):467–508, 2015.

J. Schumi and J. T. Wittes. Through the looking glass: understanding non-inferiority. *Trials*, 12(1):1–12, 2011.

D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison. Hidden technical debt in machine learning systems. In *Neural Information Processing Systems (NeurIPS)*, pages 2503–2511, 2015.

M. P. Sendak, J. D'Arcy, S. Kashyap, M. Gao, M. Nichols, K. Corey, W. Ratliff, and S. Balu. A path for translation of machine learning products into healthcare delivery. *European Medical Journal Innovations*, 10:19–00172, 2020.

V. Shankar, R. Roelofs, H. Mania, A. Fang, B. Recht, and L. Schmidt. Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning (ICML)*, 2020.

E. W. Steyerberg and F. E. Harrell. Prediction models need appropriate internal, internal–external, and external validation. *Journal of Clinical Epidemiology*, 69:245–247, 2016.

H. Suresh and J. V. Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2019.

D. Szucs and J. P. Ioannidis. Sample size evolution in neuroimaging research: an evaluation of highly-cited studies (1990-2012) and of latest practices (2017-2018) in high-impact journals. *NeuroImage*, page 117164, 2020.

T. Tasdizen, M. Sajjadi, M. Javanmardi, and N. Ramesh. Improving the robustness of convolutional networks to appearance variability in biomedical images. In *International Symposium on Biomedical Imaging (ISBI)*, pages 549–553. IEEE, 2018.

R. Tatman, J. VanderPlas, and S. Dane. A practical taxonomy of reproducibility for machine learning research. In *ICML workshop on Reproducibility in Machine Learning*, 2018.

D. Teney, K. Kafle, R. Shrestha, E. Abbasnejad, C. Kanan, and A. v. d. Hengel. On the value of out-of-distribution testing: An example of goodhart's law. In *Neural Information Processing Systems (NeurIPS)*, 2020.

R. Thomas and D. Uminsky. The problem with metrics is a fundamental problem for ai. *arXiv preprint arXiv:2002.08512*, 2020.

W. H. Thompson, J. Wright, P. G. Bissett, and R. A. Poldrack. Meta-research: Dataset decay and the problem of sequential analyses on open datasets. *eLife*, 9:e53498, 2020.

E. J. Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1): 44–56, 2019.

A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1521–1528, 2011.

G. Varoquaux. Cross-validation failure: small sample sizes lead to large error bars. *Neuroimage*, 180:68–77, 2018.

M. Voets, K. Møllersen, and L. A. Bongo. Replication study: Development and validation of deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *arXiv preprint arXiv:1803.04337*, 2018.

C. Von Bearnensquash. Paper gestalt. *Secret Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2010.

C. Wachinger, B. G. Becker, and A. Rieckmann. Detect, quantify, and incorporate dataset bias: A neuroimaging analysis on 12,207 individuals. *arXiv preprint arXiv:1804.10764*, 2018.

K. L. Wagstaff. Machine learning that matters. In *International Conference on Machine Learning (ICML)*, pages 529–536, 2012.

J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, J. Samper-González, A. Routier, S. Bottani, D. Dormont, S. Durrleman, N. Burgos, O. Colliot, et al. Convolutional neural networks for classification of alzheimer's disease: Overview and reproducible evaluation. *Medical Image Analysis*, page 101694, 2020.

J. Wilkinson, K. F. Arnold, E. J. Murray, M. van Smeden, K. Carr, R. Sippy, M. de Kamps, A. Beam, S. Konigorski, C. Lippert, et al. Time to reality check the promises of machine learning-powered precision medicine. *The Lancet Digital Health*, 2020.

M. J. Willemink, W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R. M. Summers, D. L. Rubin, and M. P. Lungren. Preparing medical imaging data for machine learning. *Radiology*, page 192224, 2020.

J. K. Winkler, C. Fink, F. Toberer, A. Enk, T. Deinlein, R. Hofmann-Wellenhof, L. Thomas, A. Lallas, A. Blum, W. Stolz, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatology*, 155(10):1135–1141, 2019.

C.-W. Woo, L. J. Chang, M. A. Lindquist, and T. D. Wager.

Building better biomarkers: brain models in translational neuroimaging. *Nature Neuroscience*, 20(3):365, 2017.

X. Yu, H. Zheng, C. Liu, Y. Huang, and X. Ding. Classify epithelium-stroma in histopathological images based on deep transferable network. *Journal of Microscopy*, 271(2): 164–173, 2018.

J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Medicine*, 15 (11):e1002683, 2018.

O. Zendel, M. Murschitz, M. Humenberger, and W. Herzner. How good is my test data? introducing safety analysis for computer vision. *International Journal of Computer Vision*, 125(1-3):95–109, 2017.

S. K. Zhou, H. Greenspan, C. Davatzikos, J. S. Duncan, B. van Ginneken, A. Madabhushi, J. L. Prince, D. Rueckert, and R. M. Summers. A review of deep learning in medical imaging: Image traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, pages 1–19, 2020.