

The Dark Side of Radiomics: On the Paramount Importance of Publishing Negative Results

Irène Buvat and Fanny Orlhac

Imagerie Moléculaire In Vivo, CEA-SHFJ, INSERM, CNRS, Université Paris-Sud, Université Paris-Saclay, Orsay, France

See the associated article on page 1545.

Over the past few years, we have witnessed an exponential increase in the number of radiomic-related publications. In the PET literature, a PubMed search using a criterion of “(radiomic OR radiomics OR texture OR textural) AND PET” yielded 37 publications in 2015 increasing to 110 in 2018. Interestingly, an extensive survey of these publications demonstrates that 94% reported positive or “promising” results involving some sophisticated radiomic features, with large variations in the performance supporting that conclusion. As few as 6% of them clearly concluded negative results, including the paper by Saadani et al. (1) published in this issue of *The Journal of Nuclear Medicine*. These numbers highlight a publication bias well acknowledged in many different research fields, with an ongoing debate about the actual false-discovery rate in the medical literature (2–4). This publication bias has well-understood roots. From the editors’ standpoint, positive results are often thought to be much more exciting and valuable than negative ones and are more likely to be cited, hence favorably increasing the journal influence metrics (e.g., 2-y impact factor) (5). From the authors’ perspective, positive results are more rewarding than negative findings and might better contribute to boosting their careers. As a consequence, negative radiomic results, the dark side of radiomics, currently remain mostly unpublished.

Yet, publishing negative results is a must for ethical reasons. Criteria for publishing should be the quality of the study and its statistical power, whatever the outcome. A clinically or biologically relevant question and a methodologically well-designed study should warrant publication, no matter if the null hypothesis is rejected or not. In some instances, negative studies might be even more impactful than positive ones, as they may challenge existing paradigms and invite investigators to focus research efforts on different paths. The conclusion of the Saadani et al. paper (1) is that although B-rapidly accelerated fibrosarcoma (BRAF) mutation drives the mitogen-activated protein kinase pathway and glucose metabolism in some cancers, BRAF valine 600 status could not be successfully detected using radiomic fea-

tures calculated from ^{18}F -FDG PET/CT in melanoma patients. This observation should be an incentive to further explore the connection between the genetic mutations or pathway alterations and their macroscopic consequences detectable using our in vivo imaging devices. It is reasonable to expect that some biologic dysfunctions resulting from mutations will produce a cascade of events that might ultimately yield a signal detectable by our exquisite molecular imaging scanners. However, both the magnitude and the spatiotemporal extent of the biologic effect will determine our ability to identify an abnormality from in vivo images using a given radiotracer. Investigations of the relationships between the triggering signal (here a mutation) and the net observable result (here a change in tumor glucose metabolism) are absolutely needed for two reasons: first, to establish realistic expectations regarding the potential power of radiomic features; second, to use radiomic observations as a driver to formulate more precise biologic assumptions regarding the underlying processes and subsequently test them. Advancing that field will require the publication of both positive and negative radiomic results.

One could argue that given the overwhelming number of radiomic-related publications reporting positive results, which might be the trees that hide the forest, the publication of negative results in that domain will be practically challenging. What we need are methodologically sound, properly powered, and robust radiomic studies addressing a biologically driven hypothesis and described in such a way that independent investigators can reproduce and confirm the findings. Indeed, replication has often more scientific value than original discovery in the radiomic field. To the best of our knowledge, since 2014 no PET radiomic model has ever been validated by an independent group. Even worse, misleading interpretation of some of the most famous radiomic findings (6) has recently been demonstrated (7), confirming that some radiomic features initially interpreted as biomarkers of tumor heterogeneity were actually surrogates of tumor volume, as shown as early as 2014 (8,9). These findings suggest that radiomic-related publications should be thoroughly designed with sufficient statistical power, described in such a way that they can be repeated, and they should also include well-supported interpretation, possibly based on dedicated experiments.

Among the methodologic arsenal that can be used to chase spurious and confounding effects in radiomic studies, “sham” data and permutation/randomization tests could be used more often. Sham data can be obtained from healthy regions or by reshuffling samples in artificial groups or randomly changing voxel values (7). Such sham data are useful to check that the findings are present in the real data and not in the sham data. It is thanks to such sham data that the erroneous interpretation of the radiomic features highlighted in one study (6) could be nicely demonstrated in

Received Aug. 20, 2019; revision accepted Sep. 10, 2019.
For correspondence or reprints contact: Irène Buvat, IMIV, CEA-SHFJ, INSERM, CNRS, Université Paris-Sud, Université Paris Saclay, 4 place du Général Leclerc, 91400 Orsay, France.
E-mail: irene.buvat@u-psud.fr
Published online Sep. 20, 2019.
COPYRIGHT © 2019 by the Society of Nuclear Medicine and Molecular Imaging.
DOI: 10.2967/jnumed.119.235325

another (7). Permutation/randomization tests also yield a precise estimate of the distribution of a test statistic under the null hypothesis in a nonparametric setting by calculating all possible values of the test statistic under rearrangements of the labels on the observed data points. They are frequently used in genomic data analysis to assign a statistical significance to the findings (10). Last, statistical reviewing should be systematic, as it is effective in improving the quality of published articles.

The validity and soundness of radiomic studies also depend on how the “vibration of effects” is handled. Vibration of effects means that results can differ over a broad range depending on how the analysis is performed. This is observed when many variations of analyses are used, such as by selecting different variables, making statistical adjustments, testing different statistical models, sorting the data differently, or using different inclusion and exclusion criteria. The variability of results as a function of the chosen analysis strategy gives an indication about their robustness. Reporting only the analyses that yield a particular result, called the chrysalis effect (11), can yield a bias, whereas consistent conclusions across various methodologic approaches, as reported in the Saadani et al. paper (1), provide a persuasive indication of the correctness of the findings.

Radiomic results should be interpreted with great circumspection. Radiomic observations are easy to report given the wide availability of software enabling the calculation of radiomic features and of packages supporting sophisticated statistical analysis and data mining. Yet, radiomics should now go beyond reporting associations and designing predictive models. It should explain what the observed patterns and associations mean from a biologic point of view, so as to enhance our understanding of the mechanisms. Dedicated experiments and multiomics approaches are needed to fully take advantage of the radiomic approach, so that the radiomic phenotype and associated models can be additional exploratory tools to untangle the complexity of the biologic processes and of their macroscopic repercussions.

A more stringent selection of radiomic studies should thus be based on tougher editorial standards. More skepticism on the part of referees is needed to avoid the inflation of false-positive findings or overzealous interpretation of radiomic results. In that respect, the 5 aspects discussed above might be systematically considered: biologic rationale justifying the study, ability to independently replicate and reproduce the findings, comprehensive control of false discoveries validated by systematic statistical reviewing, investigation of the vibration of the effects, and proofs supporting the interpretation of the results. Careful selection of publishable manuscripts based on these criteria will then leave room for publishing radiomic studies yielding negative results and confirmatory radiomic studies, both of which are absolutely needed. Dedicated sections in top-tier journals could actually be devoted to important negative and confirmatory studies. Still, it will remain virtually impossible to publish all negative radiomic

findings that would be useful to avoid unnecessary and costly repeats of experiments already done. Several practical solutions can be suggested. First, proposed future studies could be registered in international radiomic trial registries, so that it would be easier to determine whether a given question has already been or is being addressed by any investigator before an investment is made in that research. Second, negative results could be made accessible in public repositories, such as arxiv.org, and referred to in review papers to facilitate their identification. Last, funding agencies should request a registration identification number of the funded radiomic studies and access to the results, including the negative findings. All these actions together could considerably reduce the number of weekly contributive radiomic studies, clarify the state of the art, balance positive against negative radiomic results, and contribute to a proper identification of radiomic models that are useful either for patient care or for advancing our knowledge regarding how in vivo imaging can probe microscopic biologic mechanisms. In addition, establishing and enforcing best practices should help us get continuous support from external stakeholders that provide funding to advance sustainable research in radiomics.

DISCLOSURE

No potential conflict of interest relevant to this article was reported.

REFERENCES

1. Saadani H, van der Hiel B, Aalbersberg EA, et al. Metabolic biomarker-based BRAFV600 mutation association and prediction in melanoma. *J Nucl Med.* 2019;60:1545–1552.
2. Ioannidis JPA. Why most published research findings are false. *PLoS Med.* 2005; 2:e124.
3. Jager LR, Leek JT. An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics.* 2014;15:1–12.
4. Chalkidou A, O’Doherty MJ, Marsden PK. False discovery rates in PET and CT studies with texture features: a systematic review. *PLoS One.* 2015;10:e0124165.
5. Duyx B, Urlings MJE, Swaen GHM, Bouter LM, Zeegers MP. Scientific citations favor positive results: a systematic review and meta-analysis. *J Clin Epidemiol.* 2017;88:92–101.
6. Aerts HJWL, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun.* 2014; 5:4006.
7. Welch ML, McIntosh C, Haibe-Kains B, et al. Vulnerabilities of radiomic signature development: the need for safeguards. *Radiother Oncol.* 2019;130:2–9.
8. Orhac F, Soussan M, Maisonobe JA, Garcia CA, Vanderlinden B, Buvat I. Tumor texture analysis in ¹⁸F-FDG PET: relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis. *J Nucl Med.* 2014;55:414–422.
9. Buvat I, Orhac F, Soussan M. Tumor texture analysis in PET: where do we stand? *J Nucl Med.* 2015;56:1642–1644.
10. Dudoit S, Popper Shaffer J, Boldrick JC. Multiple hypothesis testing in microarray experiments. *Stat Sci.* 2003;1:71–103.
11. O’Boyle EH, Banks GC, Gonzales-Mulé E. The chrysalis effect: how ugly initial results metamorphosize into beautiful articles. *J Manage.* 2014;43: 376–399.