Version of Record:

Mazer, B.L., Homer, R.J. & Rimm, D.L. False-positive pathology: improving reproducibility with the next generation of pathologists. Lab Invest 99, 1260–1265 (2019) doi:10.1038/s41374-019-0257-2

https://www.nature.com/articles/s41374-019-0257-2

False-positive pathology: Improving reproducibility with the next generation of pathologists

Benjamin L. Mazer, Robert J. Homer, and David L. Rimm

Corresponding Author:

Benjamin L. Mazer, MD, MBA

Department of Pathology, Yale University School of Medicine, New Haven, CT

Email: benjamin.mazer@yale.edu

Address: 310 Cedar Street LH 108, PO Box 208023, New Haven, CT 06520

Robert J. Homer, MD, PhD

Department of Pathology, Yale University School of Medicine, New Haven, CT

Email: robert.homer@yale.edu

David L. Rimm, MD, PhD

Department of Pathology, Yale University School of Medicine, New Haven, CT

Email: david.rimm@yale.edu

Abstract

The external validity of the scientific literature has recently come into question, popularly referred to as the “reproducibility crisis.” It is now generally acknowledged that too many false positive or non-reproducible results are being published throughout the biomedical and social science literature due to misaligned incentives and poor methodology. Pathology is likely no exception to this problem, and may be especially prone to false positives due to common observational methodologies used in our research. Spurious findings in pathology contribute inefficiency to the scientific literature and detrimentally influence patient care. In particular, false positives in pathology affect patients through biomarker development, prognostic classification, and cancer overdiagnosis. We discuss possible sources of non-reproducible pathology studies and describe practical ways our field can improve research habits, especially among trainees.

The reproducibility crisis

The external validity of the scientific literature has recently come under question, as summarized by John Ioannidis' suggestion that "most published research findings are false."[1] This problem is popularly referred to as the "reproducibility crisis."[2,3] While scientific findings that are later disproven or fail to replicate are part and parcel of the scientific process, commonly used statistical methods are intended to limit the number of "false positives" identified by researchers. However, there is concern these statistical "checks and balances" are subverted even by well-intentioned researchers and in the absence of outright fraud.

Over-reliance on simple statistical tools such as null hypothesis significance testing allows researchers to either consciously or unconsciously alter research methods in order to achieve results at a pre-defined level of statistical significance (typically $p<0.05$). These methods are collectively referred to as "p-hacking."[4] This behavior is incentivized because non-significant results or simple replications of prior studies are subject to publication bias, meaning a disproportionate number of "positive" results make it into the published literature while non-significant results remain unpublished or published in less widely-read journals. The bias toward "positive" results occurs at different steps in the research process; for example, when researchers choose not to submit non-significant results ("file drawer effect"[5]) or through editor or reviewer rejection of non-significant results. This problem is likely getting worse, as evidenced by the declining number of "negative" findings in the scientific literature as a whole, even as major discoveries become harder to achieve.[6]

The misaligned incentives and unreliable methods contributing to non-reproducible scientific results influence many disciplines: biomedical research[7], psychology[8], economics[9], and

others[10]. It is likely that the field of pathology suffers from similar problems, yet little attention has been paid to it. In this perspective, we discuss our concern that pathology may be especially prone to false positives. Further, these non-reproducible methods are unintentionally being transmitted to the next generation of pathologists during graduate medical education, incentivized by mentor and trainee desire for advancement through quantity, rather than the quality, of contributions. We hope to begin a conversation on how better practices can realistically be encouraged.

False positives in pathology

The "false positive" problem is neither theoretical nor foreign to pathologists. These scenarios may be frustrating to individual pathologists, but they also suggest consequences for patient care, with potentially inaccurate results being used to drive diagnoses and treatment. Most pathologists should have experienced the "vicious cycle" of biomarker studies.[11, 12] An initial small, single-institution study of an immunohistochemical stain or molecular genetic test is published suggesting the new marker has exquisite sensitivity and specificity for a given diagnosis. Follow up studies from other institutions are later published showing conflicting results. Only years later is a larger, possibly multi-center study published providing a more accurate answer as to the true sensitivity and specificity of the marker, often showing the biomarker is far less valuable than originally described.

Another example of false positives in the pathology literature is the proliferation of "independent prognostic markers" for cancers based on observational studies. Taking the literature as a whole, one would be led to believe that nearly every observable histologic feature

or genetic variant influences patient outcome. Some pathologic markers will have great prognostic value, but it is unlikely so many variables simultaneously, independently, and measurably influence the patient's disease course. This paradox is similar to the observation that nearly all foods have been shown to influence mortality in nutritional epidemiology studies.[13] In truth, many of these findings are not valid, and few are truly independent predictors not influenced by other biological and clinical variables. These results are instead due to the small, underpowered or biased nature of most observational studies. Unfortunately, the field of diagnostic pathology sometimes adopts prognostic markers or sub-classifications into formal guidelines, such as those by the World Health Organization or American Joint Committee on Cancer, before rigorous reproducibility studies have been performed.[14]

Among the most influential "false positives" in pathology are those that result in widespread cancer overdiagnosis – the diagnosis of cancer in a patient whose tumor was not destined to cause harm.[15, 16] Descriptive histopathology of malignancies has traditionally derived primarily from small studies of symptomatic patients at high risk of negative outcome. These studies, even if rigorous, will be non-representative when applied to an unselected screening population. For example, an early, seminal study[17] of the follicular variant of papillary thyroid carcinoma (PTC) relied on only six cases, all of them symptomatic as goiter or lymphadenopathy. This work and others which extended it elevated characteristic nuclear features as the defining histology of PTC. After the widespread adoption of diagnostic ultrasound technology, the incidence of small, asymptomatic, and indolent thyroid "cancers" increased, causing an epidemic of overdiagnosis without improvement in population-level mortality.[18] This epidemic is caused primarily by increasing diagnosis of PTC, especially the follicular variant,

which until recently still relied on these same nuclear features.[19] This pattern has been repeated in a variety of pre-malignant and malignant lesions.

Discipline-specific vulnerability to false positives

A recent study has empirically evaluated statistical quality in diagnostic pathology research.[20] Bahar et al. found the majority of cytopathology studies examined were retrospective, and statistical errors and omissions were common. Drawing on the literature from other disciplines, there are a number of ways the diagnostic pathology literature may be particularly prone to non-reproducible methods.

First, pathologists are often not comfortable with statistical methods.[21] Pathologists report receiving very little formal training during graduate medical education and self-report a lack of knowledge of even common methods. In this context, emerging approaches to address false positives, such as Bayesian analysis[22] and causal inference[23], will remain underutilized.

Second, clinically-oriented pathology studies nearly universally rely on observational – and often subjective – methods[24], which are notoriously vulnerable to our ability to select and manipulate study variables, interpretive thresholds, and controlling factors, allowing for a multitude of potential results. This abundance of "researcher degrees of freedom"[25] is bound to produce a statistically significant result if that is what is sought, particularly when corrections for multiple hypothesis testing are not performed. The ability to completely alter an observational result based solely on modifying the analysis has been termed the "vibration of effects."[26]

The randomized controlled trial (RCT) is the traditional solution to the limitations of observational studies, yet implementation of RCTs in pathology faces practical barriers. Randomizing patients to different diagnostic approaches is challenging for histologic diagnosis, which is still based on individual judgement and experience. Changes in clinical outcome from diagnostic variation are also challenging to measure in a controlled setting, as diagnosis is an intermediary step between illness and outcome, resulting in a cascade of diverse treatment responses to a given pathologic diagnosis. A recent RCT assessing whether the diagnostic impact of high sensitivity troponin translates to improved clinical outcomes in myocardial infarction shows these trials are theoretically possible for at least automated laboratory testing, but even this remains expensive and challenging to organize.[27]

Third, due to challenges with data collection and the nature of some rare diseases, studies typically contain small sample sizes. Studies are therefore underpowered to detect modest effects or correlations while also paradoxically more likely to show false positives.[28] Worsening the situation, sample sizes can be adjusted in real time simply by collecting more years' worth of data in order to increase the chance of a (potentially spurious) statistically significant result. This seemingly benign process is part of what has led to false positive results in the field of psychology.[25]

Fourth, increasing availability of tumor registries and databases, as well as other "big data" sources, allow a pathologist to test many hypotheses with ease, only publishing ones that appear "statistically significant."[29] Genetic association studies, for example, have historically been prone to this manner of false positives.[30] While this process of testing multiple hypotheses on a fixed data set intuitively feels productive, its tendency to produce false positive p-values and disingenuous scientific logic is sometimes derisively referred to as "data-dredging" or

"hypothesizing after the results are known" (HARKing).[31] While hypothesis-generating discovery work can sometimes find interesting observations using these methods, they should be recognized as "discovery science" and published only after extensive validation.

Fifth, trainees are incentivized to increase the quantity of research findings at the expense of rigor, as presentation of research abstracts or publications are often tied to funding of conference attendance, residency or fellowship promotion, and other forms of career advancement. Trainees entering academic pathology continue to be subject to these "publish or perish" pressures.

Creating new research habits

Addressing these misaligned incentives is more challenging than identifying them. Nevertheless, pathology should join other fields in biomedicine and throughout science in openly discussing this problem[32], with a focus on how non-reproducible habits may not be transmitted to future generations of pathologists. We make the following suggestions:

Encourage trainees to clarify their analytical plan before any data are collected. Starting with a prescribed scientific hypothesis and a prospective statistically-powered approach is challenging, but can help prevent false discoveries. If this process is rigorously adhered to, it naturally reduces inappropriate statistical methods, p-hacking, and HARKing. Consultation with professionally-trained statisticians is integral for this process, and when this has been done it should be clearly documented within a manuscript. Non-punitive incentives could be utilized to encourage such behavior. Pathology publications and conferences, for example, could display distinctive levels of evidence "badges" to indicate when a study reports a prospective analytical

plan and power analysis (**Figure 1**). Publication "badges" to encourage reproducibility have been successfully implemented in other disciplines.[32]

Encourage trainees to pursue multi-institution studies. Single-institution studies are vulnerable to underpowered samples, unique patient populations, and idiosyncratic diagnostic methods, all of which increase the rate of false positives. Even a well-designed study can produce a false positive if the underlying biological and environmental aspects of patient populations differ across institutions. Single institution studies may also mask problems with inter-observer reproducibility and the challenges of implementing a new diagnostic approach in different practice contexts. These methodologic inconsistencies also reduce our field's ability to apply meta-analysis as a means to synthesize disparate studies with conflicting results, leading to an underutilization of meta-analysis throughout the discipline.[33] With the convenience of internet-mediated communication, multi-institutional approaches should always be considered first. Trainees in particular may not only be more comfortable with digital collaboration but even relish the opportunity to interact with trainees from other institutions, especially when connected by their mentors who already have extramural relationships.

Do not tie conference participation or funding to trainee abstract submission. Conferences provide a wide range of benefits to trainees, including education, leadership, networking, and even potentially reducing burnout through camaraderie. By encouraging residents to submit research abstracts even when residents are not truly engaged with this process or when results are premature will lead to false positives. While requiring abstracts as condition for meeting attendance has been an incentive for research, it is often an incentive for poor quality research in order to get a "ticket" to the meeting.

Encourage collaboration with experienced mentors and across disciplines. Too often in surgical pathology a pathologist comes up with a musing or an idea, and this becomes the nidus for a "resident project" without further input by senior pathologists and statistical experts. Consulting a senior pathologist with many years of experience may lend wisdom to the process and even prevent unnecessary use of resources on hypotheses with low prior probability or few practical implications. Consulting mentors across disciplines my yield similar wisdom, where a statistician may advise on numbers required to prove the hypothesis or a basic or translational scientist may propose a mechanistic experiment that would add rigor and reliability to the proposed study.

Challenges to improving reproducibility

The most salient barrier to implementing these changes is the pressure on early-career pathologists to gain academic currency through both quantity of publications and first authorship, both of which become harder with fewer, more collaborative studies.[34] More challenging but perhaps more influential changes would entail pathologists' receiving "academic credit" for activities like contributing to a shared study dataset, publishing pre-specified analyses ("registered reports"), providing open data sets, and performing replicated studies.[32] This can be encouraged by rewarding middle author participation in higher impact work in formal evaluations of junior colleagues.

We have attempted to provide recommendations that do not require large increases in research funding, but it should not be ignored that many of the replication problems in diagnostic pathology are exacerbated by a general lack of funding for diagnostic medicine. A recent

National Academy of Medicine report, for example, confirmed this low priority, declaring that "available research funding for diagnosis often targets specific diseases but not diagnosis as a whole or the diagnosis of several diseases with similar presentations. Diagnosis and diagnostic error are not a focus of federal health services' research efforts."[35]

Conclusion

Pathology is likely as vulnerable to false positive studies as other scientific disciplines, if not more so. Some of the causes of non-reproducible research are due to well-known but entrenched obstacles: lack of research funding, barriers to collaboration, lack of professional statistical assistance, and long-standing biases in the scholarly publishing system. Other false positives are the unintentional result of poor knowledge of effective research habits by trainees and mentors alike. The advice in this essay is intended for but not limited to trainees. While trainees are hopefully open-minded when it comes to improving their skills in practice and research, senior and mid-career pathologists also need to consider the issues raised in this work when serving as mentors to these trainees.

Some experts have proposed addressing this problem by lowering the threshold for a "statistically significant" p-value, perhaps to as low as 0.005.[36] While this suggestion does crudely reduce the chance of false positives (with the tradeoff of increased false negatives), it does not correct the underlying biases causing the non-reproducibility problem. Therefore, other changes such as those suggested in this perspective will still be required. Some journals have recently banned p-values entirely in favor of emphasizing effect sizes and confidence intervals.[23]

False positives in pathology are neither a nuisance nor a fluke. They influence patient care, waste valuable resources, and clog the machinery of scholarly publishing. This perspective serves only as an outline of the problems the field faces and presents some potential ways to address them. Solving these problems will require communication with all stakeholders, as well as deliberate, incremental changes to pathology's research infrastructure. We suggest advocating for our trainees is a good first step.

Acknowledgements. The authors thank Burak Bahar, MD for his feedback while preparing this manuscript.

Conflict of interest disclosure.

David Rimm discloses the following: He serves as a consultant, advisor and/or serves on a scientific advisory board for Amgen, Astra Zeneca, Agendia, Biocept, BMS, Cell Signaling Technology, Cepheid, Daiichi Sankyo, GSK, InVicro/Konica Minolta, Merck, NanoString, Perkin Elmer, PAIGE.AI, and Ultivue. He holds equity in PixelGear. His research lab has funding from Astra Zeneca, Cepheid, Navigate/Novartis, NextCure, Lilly, Ultivue, and Perkin Elmer/Akoya fund. Benjamin Mazer and Robert Homer have no conflicts of interest to declare.

References

1 Ioannidis JP. Why most published research findings are false. *PLoS Med* 2005;2:e124.

2 Bishop DVM. What is the reproducibility crisis in science and what can we do about it? *CBU Open Science Workshop*. Cambridge, UK, 2016. Available online: http://www.mrc-cbu.cam.ac.uk/wp-content/uploads/2016/09/Bishop_CBUOpenScience_November2016.pdf

3 Baker M. 1,500 scientists lift the lid on reproducibility. *Nature* 2016;533:452-454.

4 Head ML, Holman L, Lanfear R, *et al.* The extent and consequences of p-hacking in science. *PLoS Biol* 2015;13:e1002106.

5 Rosenthal R. The file drawer problem and tolerance for null results. *Psychol Bull* 1979;86:638-641.

6 Fanelli D. Negative results are disappearing from most disciplines and countries. *Scientometrics* 2012;90:891-904.

7 Collins FS, Tabak LA. Policy: NIH plans to enhance reproducibility. *Nature* 2014;505:612-613.

8 Open Science C. Estimating the reproducibility of psychological science. *Science* 2015;349:aac4716.

9 Camerer CF, Dreber A, Forsell E, *et al.* Evaluating replicability of laboratory experiments in economics. *Science* 2016;351:1433-1436.

10 Fanelli D. "Positive" results increase down the Hierarchy of the Sciences. *PLoS One* 2010;5:e10068.

11 Hayes DF, Allen J, Compton C, *et al.* Breaking a vicious cycle. *Sci Transl Med* 2013;5:196cm196.

12 Kern SE. Why your new cancer biomarker may never work: recurrent patterns and remarkable diversity in biomarker failures. *Cancer Res* 2012;72:6097-6101.

13 Ioannidis JPA. The challenge of reforming nutritional epidemiologic research. *JAMA* 2018;320:969-970.

14 Lee LH, Yantiss RK, Sadot E, *et al.* Diagnosing colorectal medullary carcinoma: interobserver variability and clinicopathological implications. *Hum Pathol* 2017;62:74-82.

15 Welch HG, Black WC. Overdiagnosis in cancer. *J Natl Cancer Inst* 2010;102:605-613.

16 Foucar E. Diagnostic precision and accuracy in interpretation of specimens from cancer screening programs. *Semin Diagn Pathol* 2005;22:147-155.

17 Chen KT, Rosai J. Follicular variant of thyroid papillary carcinoma: a clinicopathologic study of six cases. *Am J Surg Pathol* 1977;1:123-130.

18 Davies L, Welch HG. Current thyroid cancer trends in the United States. *JAMA Otolaryngol Head Neck Surg* 2014;140:317-322.

19 Nikiforov YE, Seethala RR, Tallini G, *et al.* Nomenclature revision for encapsulated follicular variant of papillary thyroid carcinoma: A paradigm shift to reduce overtreatment of indolent tumors. *JAMA Oncol* 2016;2:1023-1029.

20 Bahar B, Pambuccian SE, Barkan GA, *et al.* The use and misuse of statistical methods in cytopathology studies: Review of 6 journals. *Lab Med* 2019;50:8-15.

21 Schmidt RL, Chute DJ, Colbert-Getz JM, *et al.* Statistical literacy among academic pathologists: A survey study to gauge knowledge of frequently used statistical tests among trainees and faculty. *Arch Pathol Lab Med* 2017;141:279-287.

22 Held L, Ott M. On p-values and Bayes factors. *Annu Rev Stat Appl* 2018;5:393-419.

23 Lederer DJ, Bell SC, Branson RD, *et al.* Control of confounding and reporting of results in causal inference studies. Guidance for authors from editors of respiratory, sleep, and critical care journals. *Ann Am Thorac Soc* 2019;16:22-28.

24 Crawford JM. Original research in pathology: judgment, or evidence-based medicine? *Lab Invest* 2007;87:104-114.

25 Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci* 2011;22:1359-1366.

26 Patel CJ, Burford B, Ioannidis JP. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *J Clin Epidemiol* 2015;68:1046-1058.

27 Shah ASV, Anand A, Strachan FE, *et al.* High-sensitivity troponin in the evaluation of patients with suspected acute coronary syndrome: a stepped-wedge, cluster-randomised controlled trial. *Lancet* 2018;392:919-928.

28 Button KS, Ioannidis JP, Mokrysz C, *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* 2013;14:365-376.

29 Gill J, Prasad V. Improving observational studies in the era of big data. *Lancet* 2018;392:716-717.

30 Qu HQ, Tien M, Polychronakos C. Statistical significance in genetic association studies. *Clin Invest Med* 2010;33:E266-270.

31 Kerr NL. HARKing: hypothesizing after the results are known. *Pers Soc Psychol Rev* 1998;2:196-217.

32 Munafò MR, Nosek BA, Bishop DVM, *et al.* A manifesto for reproducible science. *Nat Hum Behav* 2017;1:0021.

33 Kinzler M, Zhang L. Underutilization of meta-analysis in diagnostic pathology. *Arch Pathol Lab Med* 2015;139:1302-1307.

34 Poldrack RA. The costs of reproducibility. *Neuron* 2019;101:11-14.

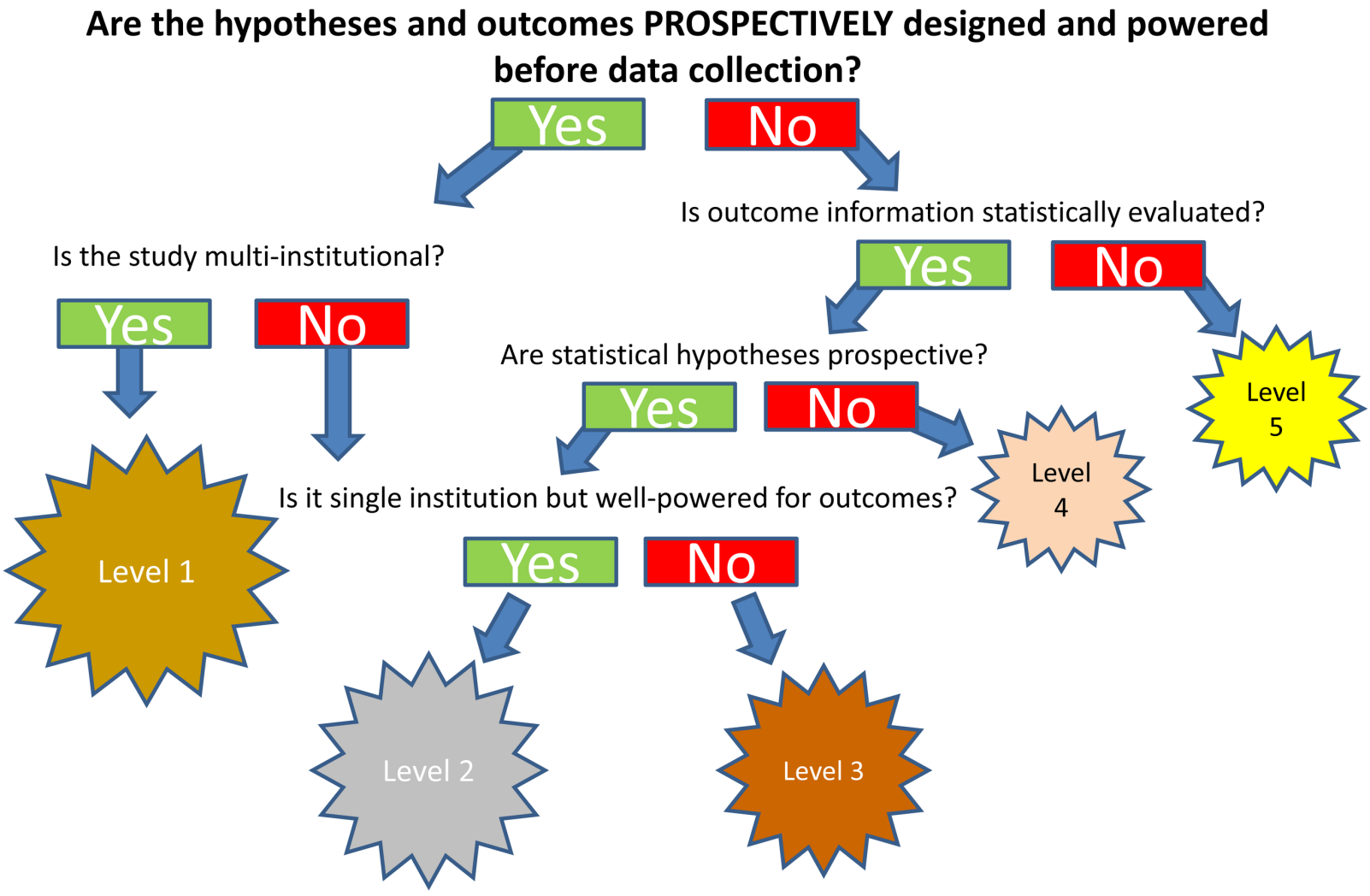
35 National Academies of Sciences, Engineering, and Medicine. Improving Diagnosis in Health Care. Washington, DC: National Academies Press, 2015, p.389.

36 Ioannidis JPA. The proposal to lower P value thresholds to .005. *JAMA* 2018;319:1429-1430.

37 Simon RM, Paik S, Hayes DF. Use of archived specimens in evaluation of prognostic and predictive biomarkers. *J Natl Cancer Inst* 2009;101:1446-1452.

Figure 1. Levels of evidence “badges” for use in publications and conferences to encourage reproducible science in pathology. These categories emphasize prospectively specifying statistical methods (no HARKing) and prospectively powering the study for the outcomes of interest to reduce false positives from underpowered samples. Levels of evidence were adapted from Simon et al.[37]

Levels of evidence "badges" for pathology studies



Levels of evidence adapted from Simon RM, Paik S, Hayes DF. JNCI. 2009;101:1446-52.