

## Getting it right matters! Covid-19 pandemic analogies to everyday life in medical sciences

What a time it has been since the world quite literally changed within weeks in early 2020. For all of us, as researchers and as normal citizens life changed drastically.<sup>1</sup> Laboratory meetings turned into lockdown video calls, cultural events were deemed dispensable and a sentimental coffee on the balcony became the luxury version of any vacation plans. Now, as the pandemic is retreating in most countries and optimistic prognosis for the upcoming months is no more mutually exclusive, people take a step back to evaluate what happened over the last months, both on a societal and a personal level.

We wish to express our deepest condolences to everyone who has lost a loved one during the pandemic, or who struggle themselves with Covid or its late consequences. A quick grasp of history tells us, that things could have turned out much, much worse.<sup>2</sup> That is not to discredit the personal tragedies caused by this pandemic, but a quest for a silver lining. Those moments of relief, of seeing positive even in the eye of severe adversity is not only helpful in preserving mental health, but it helps us make the best of the situation.

When taking a closer look, there have been many little things brought up by the pandemic, which could turn out positive in the long run. Admittedly, video calls can never supersede real, personal encounters, but questioning, whether there are things researchers (and members of any other profession) can do from home could add a lot of flexibility to the way we work. The interruption of most studies relying on human participants forced many researchers to restructure work. Taking out this cornerstone of medical science is disheartening, but then again free time to ponder can be a blessing in disguise.<sup>3-5</sup>

For contemplation to be a blessing requires worthwhile thoughts. With regard to science, the focus on Covid-19 case numbers, herd immunity, vaccinations, and appreciation of metrics for decision-making have transformed the general thought process. Laymen reflect on incidences, R-values, essential vaccination rates and might even have heard the terms specificity and sensitivity regarding antigen tests. It is even more heartening when those people start to learn from analogy and use the newly acquired scientific concepts to talk about the things they care about. Seeing everyday people engaged in science bodes well for society.

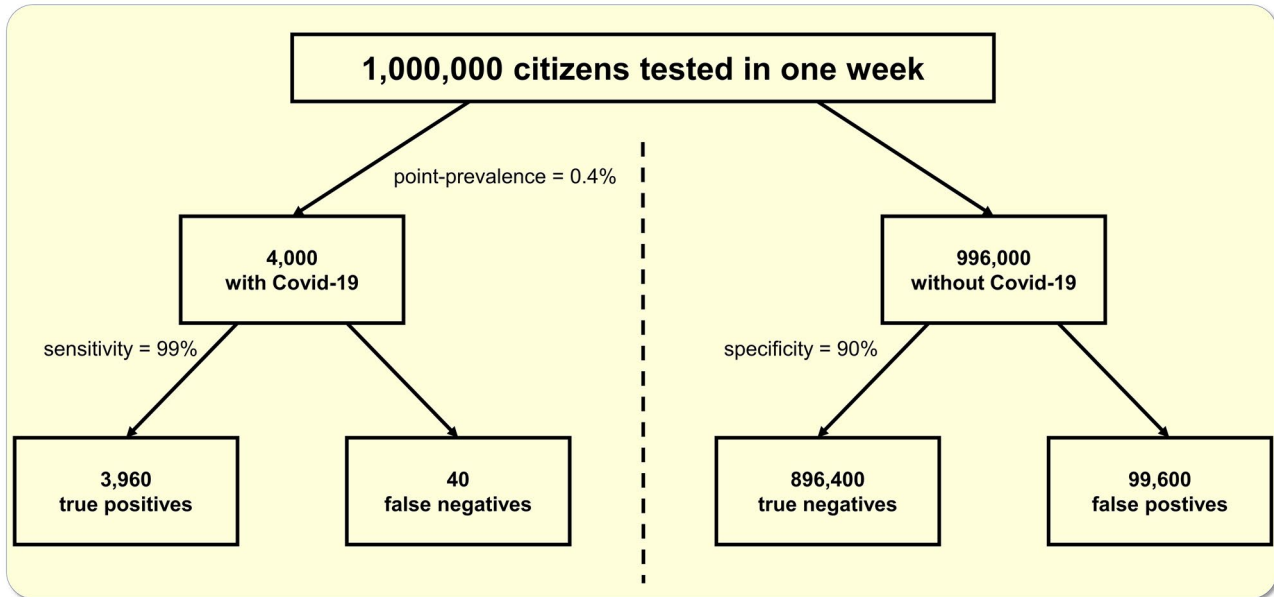
Learning sound great. Scientists quickly focussed on learning about the pandemic and the virus's characteristics.<sup>6-8</sup> They even presented first results about the virus's implication on other organ systems, beside the respiratory system.<sup>9-11</sup> Yet, is there something all scientists can learn by looking closer at the concepts Covid-19 has so plastically presented us with? It turns out there is: Because this pandemic has been so overarching, we all share common experiences that can help enhance insight into complex problems. In this sense, let us have a little thought experiment on a particular issue related to the pandemic:

Imagine taking a Covid-19 antigen test. As you are a scientist, you are aware of the concepts of specificity (true positive rate) and sensitivity (true negative rate). As the test is of very high quality (sensitivity = 0.99/specificity = 0.90), you feel very confident about the negative result you got and happily proceed with your plans (eg to visit your relatives). Now let us test your initial joy: We just assume the current Covid-19 incidence in your area to be at around 200 per 100 000 per week and your countries test rate at 1 million per week. For argument's sake, we also assume the average Covid-19 infection (mean time a patient might be contagious) to last 14 days, which leads us to a point prevalence of 0.4% or 400 per 100,000 at the time you take your test.

$$\begin{aligned} \text{Point prevalence} &= \frac{200 \text{ weekly infections}}{100\,000 \text{ inhabitants}} \\ \cdot \frac{14 \text{ days of illness}}{7 \text{ days}} &= \frac{400 \text{ presently infected}}{100\,000 \text{ inhabitants}} \end{aligned}$$

We can now plot all the information into a tree diagram and evaluate the results (Figure 1).

It turns out your feeling of relief was indeed warranted. It is very unlikely for you to be one of the false negative results and even if you had been tested positive, your chance of actually being infected would only have been around 3.8% (3960 true positives/(3960 true positives + 99 600 false positives), because of a large number of false positives. In other words: The false positive risk of your test is at about 96.2% (99 600 false positives/(3960 true positives + 99 600 false positives). This phenomenon, large numbers of false positives even



**FIGURE 1** Tree diagram portraying discovery rates in a Covid-19 antigen test paradigm. The numbers are fictional but chosen to mimic plausible real-world data. The point prevalence (percentage of Covid-19 infected people in the general public at the time of the test), the sensitivity and the specificity are provided

though test performance measures seem to be very high, has been described for many screening tests and should be considered emphatically in particular for widely advertised tests. It is of course the rationale behind secondary PCR tests for individuals tested positive: We do not want to force non-contiguous people into quarantine.

Great. After all a lot of scientists are all too familiar with PCRs, so there is not much to learn there.<sup>12-16</sup> We made some assumptions and visualized a statistical concept which most of us would admit we had not been fully considering before the pandemic. So, this is it, we can all go (or stay) home and feel satisfied, because we had a little academic workout and a better understanding of screening tests usually only cared about by a narrow circle of physicians? Not quite. One central point in learning from analogy is to have an analogy, preferably one, which relates to everyday life. Nonetheless, how can we relate basic Covid-19 test statistics with the vast diversity of your, the readers', lives in science? Workdays in science are diverse. Scientists work in a multitude of different and highly specialized fields, so finding a common ground sounds rather difficult.<sup>17-21</sup>

It is the work of David Colquhoun, which allows us to do so by presenting a beautiful analogy between screening tests and the crisis of unreproducible results in medical sciences.<sup>22,23</sup> Unreproducible results are obstacles in the way of scientific process that question scientific integrity and constrain or even terminate scientific careers.

To make the analogy, we must think about what unreproducible results are and then figure out a way to stay on top of this issue. Unreproducible results can involve inappropriate

data acquisition and handling, or false positive results. We will focus on the latter, as they are very common and readily nailed down.

To reduce the occurrence of false positive results, the scientific community has taken measures such as reporting the *P* value of statistical tests and assumed power in study design. It has been brought back to our attention recently why *P* values alone are no good representation of experimental results, as they only provide a measure of certainty without recognizing the practical relevance of a shown effect.<sup>24</sup> Moreover, keep in mind that *P* values work similarly to a test's specificity—it only provides insight into cases without a real effect, that is, only into the cases shown in the right part of the tree in Figure 1. Let us step further and think about what the *P* value and the statistical power are and why they, in their current form, may provide unreproducible results.

To do so, we will draw up another tree diagram for which we will need some assumptions: You, your working group and everyone in your institute test roughly 300 hypotheses each year altogether, which will add up to 10,500 lifetime hypotheses in 35 years in research. Then we say that in about 20% of these cases there is a real underlying effect, whether we can measure it or not. This number will most likely be much smaller, since we often explore the unknown.<sup>22</sup> We then assume everyone to design their experiment right in line with best practice, choosing an alpha level (significance level) of 0.05, a statistical power of 0.80, and all projects yield are perfectly randomized, normally distributed results, which will be tested for one hypothesis only. All these assumptions portray an ideal scenario, which will lead to the best-case number

of correctly reported and therefore most likely reproducible results.

Now for some clearing up: The *P* value does not tell us the likelihood of a shown effect being fallacious. It tells us how likely it is that we observe an effect (or a more extreme) based on chance in the particular case that there is no true underlying effect. In analogy to our antigen test example: Our chosen alpha equals the test's specificity. It tells us only how likely it is that a test is false positive. Again, keep in mind that this number is calculated based on the number of hypotheses without true effect neglecting the existence of hypotheses with true effects. Similarly, the statistical power is to be set equal to the antigen test sensitivity: The statistical power tells us how many of the true positives our test will detect. With this in mind, we can draw up our tree diagram (Figure 2).

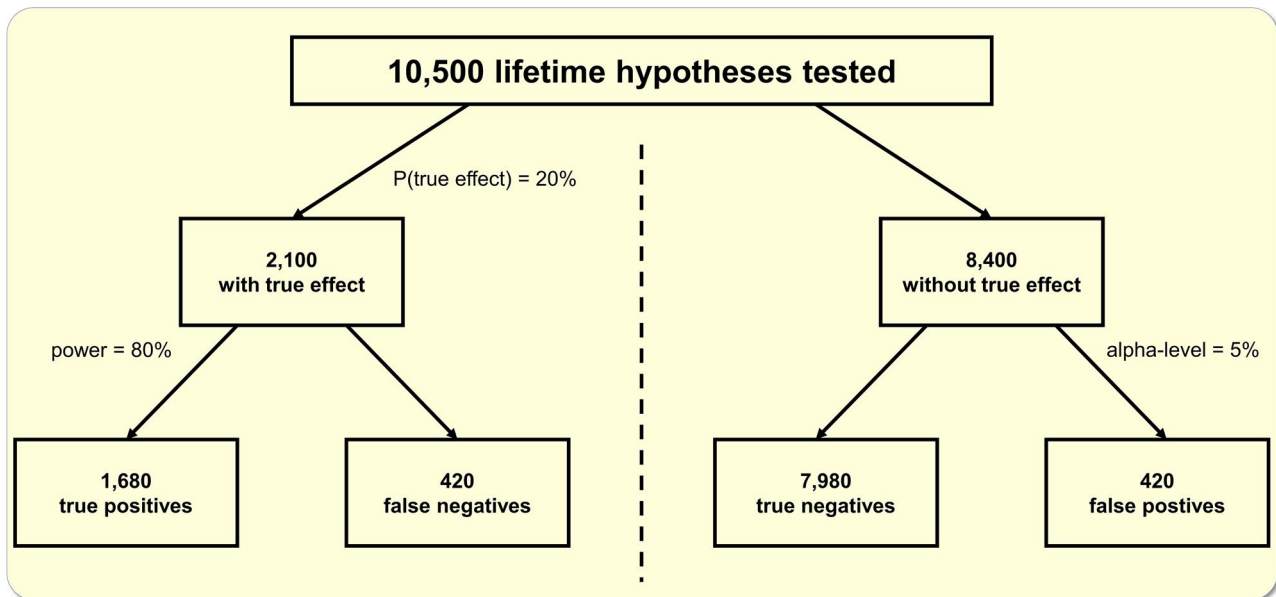
Taking a close look, we see that this best-case scenario yields grim results. If we divide the true positives by all the positives our statistical tests yielded over 35 years there are only  $1680 / (1680 + 420) = 80\%$  of all positive appearing results truly positive. Consequently, 20% of all positive (and therefore probably reported) experimental results are false discoveries. Keep in mind, this is the best-case scenario. When only slightly tweaking the numbers to a more realistic power of 60% and a true effect rate of 10% we can calculate that only 57% of all reported results are indeed genuine, leaving us with a false discovery risk of 43%. This is, still, assuming the best-case scenario of completely randomized, normally distributed, single-variable tests and the assumptions we made are not unrealistic and still more on the optimistic side. Button et al stated in a widely recognized paper that a serious estimation of the median statistical power in

published neuroscience research projects may be as low as somewhere between 8% and 31%.<sup>25</sup>

All of this is already bad enough, but it gets even worse: When accepting low statistical power, you do not only sacrifice the reliability of your results. Your reported effect sizes will inflate beyond what the actual effect is. The reason for this inflation comes quite naturally: Picture the possible measured effect sizes normally distributed around their true mean. A power of 0.3 tells you that only 30% of your experiments with a true underlying effect will show a true-positive outcome. Quite intuitively, the positive test results skew strongly towards the instances of data in which the effect appears to be large. Plainly said: If your observed effect is twice as large as the real effect, your test is way more likely to detect it as positive. This leads to reported effect sizes being overblown and, yet again, not reproducible. The lower the statistical power, the more false negative results are produced by your test, which will skew its detections more and more towards inflated effect sizes.

The situation is dire, to say the least. With our analogy in mind, we can now picture why there are so many unreproducible results published in scientific literature. So, is it time to throw your hands up and realize that we are powerless against the statistical fate of science we have just worked out? It is not.

As for most problems, its solution begins with becoming aware of it. In this case, it drives us already halfway home. Listen to Colquhoun and take some easy measures to drastically mitigate these effects.<sup>22,23</sup> First, *getting things right matters!* We must make sure that we set up our experiments properly. Statistical power analyses are not just a



**FIGURE 2** Tree diagram portraying lifetime research. The numbers are fictional but chosen to mimic plausible real-world data. The probability of a true underlying effect is provided as *P*(true effect). The statistical power and the chosen alpha level are provided in parallel to the sensitivity and specificity in Figure 1

statistician's way of passing time, they matter. A power of at least 0.8 helps tremendously to keep inflating effect sizes and false positive results in check. Second, there is the  $P$  value, and we simply should be more stringent with it. If we apply a three-sigma ( $P \leq .001$ ) instead of the current two-sigma ( $P \leq .05$ ) rule, we can shrink the risk of false positives to values  $<2\%$ , which is in fact the range the scientific community targeted as acceptable when choosing 0.05 as their alpha level of choice.<sup>22</sup>

Additionally, changing our perception of  $P$  values would go a long way. There are experimental settings which will not allow a  $P$  value beneath .001. In those cases ( $p$  close to 0.05), we have to be cautious when interpreting our results as proof of an effect. When knowledge is meagre, a significant result makes the existence of an effect more likely, nevertheless doubts remain. Such results only indicate something interesting warranting further investigation. However, for shown reasons, drawing conclusions, or even making treatment decisions based on those results is irresponsible. Only when existing knowledge provides strong evidence for a probable effect, a significant result can be interpreted with confidence that there is a true effect.<sup>26</sup>





Ultimately, reporting the false positive risk along with  $P$  values, 95% confidence intervals and observed effect sizes in relation to relevant effect sizes, can prevent us and our fellow researchers from misinterpreting scientific results. Anyone would be more than cautious to consider a study with a high false positive risk (eg 30%) as proof of an actual effect. All we need to provide for the calculation of the false positive risk is the number of biological samples in our study, the observed  $P$  value, effect size and an estimate of the prior probability of a real effect.<sup>23</sup> The first three values are easily obtained when performing any statistical test. The prior probability can only be approximated but a value of 10% seems to be reasonable. Alternatively, one may set the prior probability to 0.5 and calculate the corresponding false positive risk ( $FPR_{0.5}$ ). Notably, higher prior probabilities mean that you are studying something well established, lower prior probabilities increase the false positive risk. Thus, a prior probability of 0.5 provides a minimum estimate for a meaningful false positive risk.<sup>27</sup> There are simple and free of charge online tools for calculating the false positive risk.<sup>28</sup> It is now in our hands as members of the research community to make use of these tools and provide our readers with reliable estimates of our studies' false positive risks. In parallel to the pandemic, there is a silver lining on the horizon, at least when looking in the right direction. We can take the necessary steps to modify the scientific process sufficiently to allow warranted confidence in its results. Just as societies are striving to utilize the lessons learned to build the post-pandemic future, we as the scientists can strive to tackle the crisis of unreproducible results. All we have to do is to do what humans are so wonderfully proficient at: Trying to *get things right* and *learn from analogy*.

## ACKNOWLEDGEMENTS

We thank Laura Josefa Dippel for insightful comments and discussions.

## CONFLICTS OF INTEREST

There is no conflict of interest to declare.

Tomas L. Bothe<sup>1</sup>   
 Andreas Patzak<sup>1</sup>   
 Rudolf Schubert<sup>2</sup>   
 Niklas Pilz<sup>1</sup> 

Rudolf Schubert and Niklas Pilz are shared senior authors.

<sup>1</sup>*Institute of Vegetative Physiology, Charité –  
 Universitätsmedizin Berlin, corporate member of  
 Freie Universität Berlin and Humboldt-Universität zu  
 Berlin, Berlin, Germany*

<sup>2</sup>*Physiology, Institute of Theoretical Medicine,  
 Medical Faculty, University of Augsburg, Augsburg,  
 Germany*

## Correspondence

Tomas L. Bothe, Institute of Vegetative Physiology, Charité –  
 Universitätsmedizin Berlin, corporate member of Freie  
 Universität Berlin and Humboldt-Universität zu Berlin,  
 Berlin, Germany.

Email: tomas-lucca.bothe@charite.de

## ORCID

Tomas L. Bothe  <https://orcid.org/0000-0001-7569-4527>  
 Andreas Patzak  <https://orcid.org/0000-0002-1088-6875>  
 Rudolf Schubert  <https://orcid.org/0000-0003-1777-1461>  
 Niklas Pilz  <https://orcid.org/0000-0002-1195-8359>

## REFERENCES

1. Khedkar PH, Patzak A. SARS-CoV-2: what do we know so far? *Acta Physiol.* 2020;229(2):e13470.
2. Spreeuwenberg P, Kroneman M, Paget J. Reassessing the global mortality burden of the 1918 influenza pandemic. *Am J Epidemiol.* 2018;187(12):2561-2567.
3. Pallubinsky H, Phielix E, Dautzenberg B, et al. Passive exposure to heat improves glucose metabolism in overweight humans. *Acta Physiol.* 2020;229(4):13488.
4. Snijders T, Aussieker T, Holwerda A, Parise G, van Loon LJC, Verdijk LB. The concept of skeletal muscle memory: Evidence from animal and human studies. *Acta Physiol.* 2020;229(3):e13465.
5. Dibner C. The importance of being rhythmic: living in harmony with your body clocks. *Acta Physiol.* 2020;228(1):e13281.
6. Kache T, Mrowka R. How simulations may help us to understand the dynamics of COVID-19 spread. – visualizing non-intuitive behaviours of a pandemic (pansim.uni-jena.de). *Acta Physiol.* 2020;229(4):e13520.
7. Pontecorvi G, Bellenghi M, Ortona E, Carè A. microRNAs as new possible actors in gender disparities of Covid-19 pandemic. *Acta Physiol.* 2020;230(1):e13538.

8. Marquez A, Wysocki J, Pandit J, Batlle D. An update on ACE2 amplification and its therapeutic potential. *Acta Physiol.* 2021;231(1):e13513.
9. Steardo L, Steardo L, Zorec R, Verkhatsky A. Neuroinfection may contribute to pathophysiology and clinical manifestations of COVID-19. *Acta Physiol.* 2020;229(3):e13473.
10. Ye Q, Lai EY, Luft FC, Persson PB, Mao J. SARS-CoV-2 effects on the renin-angiotensin-aldosterone system, therapeutic implications. *Acta Physiol.* 2021;231(4):e13608.
11. Hardenberg JHB, Luft FC. Covid-19, ACE2 and the kidney. *Acta Physiol.* 2020;230(1):e13539.
12. Fan M, Chen Z, Huang Y, et al. Overexpression of the histidine triad nucleotide-binding protein 2 protects cardiac function in the adult mice after acute myocardial infarction. *Acta Physiol.* 2020;228(4):228.
13. Lu M, Qin X, Yao J, Yang Y, Zhao M, Sun L. Th17/Treg imbalance modulates rat myocardial fibrosis and heart failure by regulating LOX expression. *Acta Physiol.* 2020;230(3):e13537.
14. Ivanova AD, Filatova TS, Abramochkin D, et al. Attenuation of inward rectifier potassium current contributes to the  $\alpha$ 1-adrenergic receptor-induced proarrhythmicity in the caval vein myocardium. *Acta Physiol.* 2021;231(4):e13597.
15. Møldrup A, Lindberg MN, Galsgaard ED, Henriksen U, Dalgaard LT, Nielsen JH. Regulation of integrin  $\alpha$ 6A by lactogenic hormones in rat pancreatic  $\beta$ -cells: implications for the physiological adaptation to pregnancy. *Acta Physiol.* 2020;229(3):13454.
16. Frees A, Assersen KB, Jensen M, et al. Natriuretic peptides relax human intrarenal arteries through natriuretic peptide receptor type-A recapitulated by soluble guanylyl cyclase agonists. *Acta Physiol.* 2021;231(3):e13565.
17. Spasova K, Föhling M. D-serine—a useful biomarker for renal injury? *Acta Physiol.* 2020;230(2):e13531.
18. Rosenberger C, Föhling M. A triple sense of oxygen promotes neurovascular angiogenesis in NG2-derived cells. *Acta Physiol.* 2021;231(1):e13578.
19. Fan X, Li K, Zhu L, et al. Prolonged therapeutic effects of photoactivated adipose-derived stem cells following ischaemic injury. *Acta Physiol.* 2020;230(1):e13475.
20. Solagna F, Nogara L, Dyar KA, et al. Exercise-dependent increases in protein synthesis are accompanied by chromatin modifications and increased MRTF-SRF signalling. *Acta Physiol.* 2020;230(1):13496.
21. Sobreiro-Almeida R, Melica ME, Lasagni L, Romagnani P, Neves NM. Co-cultures of renal progenitors and endothelial cells on kidney decellularized matrices replicate the renal tubular environment in vitro. *Acta Physiol.* 2020;230(1):e13491.
22. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. *R Soc Open Sci.* 2014;1:140216.
23. Colquhoun D. The reproducibility of research and the misinterpretation of p-values. *R Soc Open Sci.* 2017;4:171085.
24. Bothe TL, Patzak A. Significant significance? *Acta Physiol.* 2021:e13665.
25. Button KS, Ioannidis JPA, Mokrysz C, et al. Power failure: Why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci.* 2013;14(5):365-376.
26. Rowe P. *Statistik für Mediziner und Pharmazeuten.* 1st ed. Weinheim: Wiley-VCH; 2012.
27. Colquhoun D. The false positive risk: a proposal concerning what to do about p-values. *Am Stat.* 2019;73(sup1):192-201.
28. Longstaff C, Colquhoun D. False Positive Risk Web Calculator. version 1.7. 2017. <http://fpr-calc.ucl.ac.uk/>. Accessed June 16, 2021.