Confronting *p-hacking*: addressing *p-value* dependence on sample size

Estibaliz Gómez-de-Mariscal[1], Alexandra Sneider[2], Hasini Jayatilaka[3], Jude M. Phillip[4], Denis Wirtz[2,5,6], and Arrate Muñoz-Barrutia[1,6,*]

[1] Bioengineering and Aerospace Engineering Department, Universidad Carlos III de Madrid, 28911 Leganés, and Instituto de Investigación Sanitaria Gregorio Marañón, 28007 Madrid, Spain

[2] Department of Chemical and Biomolecular Engineering, The Johns Hopkins University, Baltimore, Maryland 21218, USA

[3] Department of Pediatrics, Bass Center for Childhood Cancer, Stanford University, Stanford, California 94305, USA

[4] Department of Medicine, Division of Hematology and Oncology, Weill Cornell Medicine, New York 10065, USA.

[5] Department of Oncology, The Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA

[6] Co-senior author

* Correspondence and lead contact: mamunozb@ing.uc3m.es

ABSTRACT

Biomedical research has come to rely on *p-values* to determine potential translational impact. The *p-value* is routinely compared with a threshold commonly set to 0.05 to assess the significance of the null hypothesis. Whenever a large enough dataset is available, this threshold is easily reachable. This phenomenon is known as *p-hacking* and it leads to spurious conclusions. Herein, we propose a systematic and easy-to-follow protocol that models the *p-value* as an exponential function to test the existence of real statistical significance. This new approach provides a robust assessment of the null hypothesis with accurate values for the minimum data-size needed to reject it. An in-depth study of the model is carried out in both simulated and experimentally-obtained data. Simulations show that under controlled data, our assumptions are true. The results of our analysis in the experimental datasets reflect the large scope of this approach in common decision-making processes.

KEYWORDS

Biostatistics, Bioinformatics, p-hacking, Mathematical-modeling

INTRODUCTION

In the most complex scenarios, decision making is only possible when we are able to reduce intricate working conditions to a dichotomous or binary case. Statistical hypothesis testing has always supported the ability to discriminate between different events. Yet previous methods do not always provide robust results due to dependence on the size of the datasets being tested (Altman and Krzywinski, 2017; Krawczyk, 2015; Lin et al., 2013), and requires an urgent revision (Altman and Krzywinski, 2017; Amrhein et al., 2019; Leek et al., 2017).

Typically obtained from any conventional test, the "gold standard" *p-value* has long been recognized as an unreliable but popular measure of statistical significance (Altman and Krzywinski, 2017; Amrhein et al., 2019; Halsey et al., 2015; Nuzzo, 2014). The *p-value* is itself a random variable that depends on the data used; and, therefore, has a sampling distribution. A straightforward example is as follows: the *p-value* has a uniform distribution $(0,1)$ under the null hypothesis. If the null hypothesis is not trivially assessable, it remains always possible to obtain a sufficiently small *p-value* that rejects the null hypothesis by sufficiently increasing the sample size (also called *p*-hacking) (Bruns and Ioannidis, 2016; Cumming et al., 2007; Halsey et al., 2015; Krzywinski and Altman, 2013). For instance, even when comparing the mean value of two groups with identical distribution, statistically significant differences among the groups can always be found as long as a sufficiently large number of observations is available using any of the conventional statistical tests (i.e., Mann Whitney U-test (Mann and Whitney, 1947), Rank Sum test (Wilcoxon, 1945), Student's *t*-test (Student, 1908)) (Bruns and Ioannidis, 2016). Non-parametric statistical tests for two samples, such as the Kolmogorov-Smirnov test (Smirnov, 1939), also conclude with the rejection of the null hypothesis when working with sufficiently large datasets. In other words, big data can make insignificance seemingly significant by means of the classical *p-value*. Similar to the examples in (Altman and Krzywinski, 2017; Halsey et al., 2015), Fig. 3 in the Materials and Methods further illustrates the described problem.

Despite this finding, there remain many situations for which the 'dichotomy' associated with the *p-value* is necessary for decision-making (Leek et al., 2017). Designing a robust tool devoted to this task could be an inflection point in the use of statistical tests. In this work, we aim to answer the question of when can we solidly assert that bona fide differences exist between two sets of data, independent of sample size.

To introduce our method, we first show that the *p-value* can be accurately approximated through its expression as an exponential function of the sample size n, $p(n)$:

$$p(n) = a \cdot e^{-cn} \text{ where } a, c \in \mathrm{R}^{+} \qquad 1$$

In Fig. 1a, different randomly generated normal distributions are compared using the Mann-Whitney U statistical test (Mann and Whitney, 1947) to illustrate the decrease of the function $p(n)$ with the sample size. The use of the Student's *t*-test was avoided as it is well known that the *p-value* associated to the *t*-statistic has an exponential decay(Student, 1908). Technical details about the convergence of the function $p(n)$ and

evidence about Eq. 1 holding for any statistical test are given in the Materials and Methods.

Note that the *p-value* curve, the function $p(n)$, is used to compare pairs of experimental conditions; therefore, $p(n)$ is computed as the exponential fit to the probability value of multiple sample comparisons. Hence, the parameters a and c in Eq. 1 correspond to those defining the exponential fit $p(n)$. We use the Monte Carlo cross-validation (MCCV) (Xu and Liang, 2001) as the sampling strategy: two subsets of size n (one from each of the groups to be compared) are randomly sampled and compared with a statistical test. The resulting *p-value* is stored and the procedure is repeated many times. At the end of the procedure, a large set of n-dependent *p-values* is obtained and the exponential function in Eq. 1 can be fit.

Similarly to any exponential function, $p(n)$ converges to zero. The faster the function converges, the more robust the significance. When normal distributions of standard deviation one and mean value in the range [0, 3] are compared, we see that the higher the difference among experimental conditions, the faster the decay of the exponential function that approximates $p(n)$ (Fig. 1b). We observe that the parameters a and c (Eq. 1) increase proportionally with the mean value of the distribution compared with $N(0,1)$ (Fig. 1b). With this new idea in mind, a robust decision index, $\theta_{\alpha,\gamma}$, can be mathematically defined (Eq. 10 in Materials and Methods). Note that subscripts α (statistical significant threshold) and γ (regularization parameter) are omitted from now on.

Instead of comparing a single *p-value* with the ideal statistical significance threshold α (i.e., $\alpha = 0.05$ for a 95% of statistical significance), a distance δ (Eq. 9 in the Materials and Methods) is defined to compare the function $p(n)$ with α for all n values. δ measures the difference between the areas under the constant function at level α and the area under the curve $p(n)$ (Fig. 1c). The distance δ is then used to obtain the binary index θ that indicates whether $p(n)$ and the α constant are far from each other or not. If for most values of n the function $p(n)$ is smaller than α, then $\theta = 1$, which means that there is an acceptable statistical significance. However, if θ is null, the tested null hypothesis cannot be rejected.

As the exponential function is defined for all values $n \in (-\infty, +\infty)$, it is necessary to determine a range of n for which the function $p(n)$ is meaningful. The decay of $p(n)$ is concentrated in a range between $n = 0$ and a certain value of n for which $p(n) \approx 0$ (convergence of $p(n)$); so, δ should be only calculated in that range. A parameter γ is used as a regularizer to measure the point of convergence $n = n_\gamma$, such that $p(n = n_\gamma) \approx 0$ (Fig. 1c and Eq. 8 in the Materials and Methods). Small γ values imply less restrictive decisions. Nonetheless, the experimental evaluation of the method over synthetic and real data evidences $\gamma = 5e^{-06}$ to be a reasonable choice (detailed information is given in the Materials and Methods and the Supplementary Material). Note that when $p(n)$ is determined simply by the definition of the parameters a and c in Eq. 1, the minimum data size needed to observe statistically significant differences at α-level can also be provided. As $p(n)$ continuously decreases, the value of n for which $p(n)$ is always smaller than α can be calculated easily. This value is called n_α (Fig. 1c and Eq. 12 in the Materials and Methods).

RESULTS

Both the decision index θ and the minimum data size n_α provide for intuition about the veracity of the null hypothesis of the statistical test. To illustrate this, different normal distributions were compared with the Mann-Whitney U statistical test (Mann and Whitney, 1947) with an α-level of 0.05 (Table 1 in Materials and Methods). When $N(0,1)$ is compared with $N(0,1)$, $N(0.01,1)$ and $N(0.1,1)$, θ is null; so those distributions are assumed to be equal. In the remaining comparisons though, $\theta = 1$, thus there exist differences between $N(0,1)$ and $N(\mu,1)$ for $\mu \in [0.25,3]$ (Fig. 1d). Likewise, the value of n_α increases until infinity as the mean value μ decreases when $N(0,1)$ is compared with $N(\mu,1)$ for $\mu \in [0.1,3]$. Indeed, n_α cannot be determined when $N(0,1)$ is compared with $N(0,1)$ and $N(0.01,1)$, as the null hypothesis in this case is true and therefore, $p(n)$ is a constant function, which represents the uniform distribution of *p-values* (Figs. 1e and 1f, and Fig. 3 in Materials and Methods).

To prove the generality of the proposed method, we tested its different functionalities on published and non-published data from biological experiments. The first application of the method consists in discriminating between conditions; that is, to declare whether two conditions are different or not. In this case, we wanted to determine whether cancer cells cultured in 3D collagen matrices and imaged under a light microscope changed shape after administration of a chemotherapeutic drug (Taxol) (details about data collection and processing are given in the Supplementary Material). Three different groups were compared: control cells (non-treated), and cells treated with 1 nM and 50 nM Taxol respectively. Cells exposed to low concentrations of Taxol (1 nM) remained elongated (low roundness index), i.e. $\theta = 0$ for the comparison between control cells and those treated with Taxol at 1 nM. However, when the dose was increased to 50 nM Taxol, cells became circular; therefore $\theta = 1$ when comparing cells treated with 50 nM Taxol versus control cells, or cells treated with 1 nM Taxol (Fig. 2a and Table S3 in the Supplementary Material).

Secondly, we analyzed the flow cytometry data used by Khoury *et al.* (Khoury et al., 2018) to determine the transcriptional changes induced by the in vivo exposure of human eosinophils to glucocorticoids. As it was done in the previous example, the proposed method allowed us to discriminate between treated and untreated eosinophils using the entire dataset. For that, we analyzed the eosinophil surface expression of the gene CXCR4 2 h after the exposure to 20 and 200 mcg/dL of Methylprednisolone. The eosinophils belong to 6 different healthy human subjects. With the estimation of the function $p(n)$ (Eq. 1), it is possible to conclude that the exposure of eosinophils to glucocorticoids causes a differential expression of CXCR4 (Fig. 2b), i.e. $\theta = 1$ for the comparison between vehicle and eosinophils treated with 20 and 200 mcg/dL (Table S6 in the Supplementary Material). Indeed the conclusion is the same as the one made in (Khoury et al., 2018), where only the median fluorescence intensity of the data from each subject was calculated and the resulting 6 data points were compared (Fig. 2b). However, the latter approach can lead to false conclusions when the data distribution differs or when the data deviation is large.

The last use of the method consists of analyzing whether a single specific feature of the data (variable) can fully characterize the problem at hand. For instance, many different biomolecular and biophysical features of human cells were analyzed (Phillip et al., 2017) to predict cellular age in healthy humans. This is only possible if these features contain enough information about the aging of the patients. To show that, we re-analyzed a large and a small dataset with information of nuclei morphology and cell motility respectively, collected by Philip *et al.* (Phillip et al., 2017). The information of 2 year-old human cells (the youngest one) was compared with the rest of the ages. The decay of $p(n)$ in cell nuclei area and short axis length are directly related to the age of human cells. The parameter c (Eq. 1) of the orientation of the cell nuclei is null in all cases, which indicates that this measure does not contain information about aging (Fig. 2c and Table S5 in the Supplementary Material). Moreover, the estimated function $p(n)$ for the total diffusivity of the cells of 2 year-old and 3 year-old human donors shows that even if a larger dataset was given, the result will remain the same (Fig. 2d and Table S4 in the Supplementary Material). Namely, $p(n)$ does not decrease, therefore, there is strong evidence that the null hypothesis is true (i.e. $\theta = 0$, groups behave similarly). The most extreme cases given by the differences between 2 and 96 year-old human donors, can also be detected without the need of large datasets, $n_\alpha = 11$ (Fig. 2d). That is, the estimation of $p(n)$ allows one to decide whether it is valuable to collect new data to determine differences among the studied groups, or not.

DISCUSSION

The use of statistical hypothesis testing is largely extended and well established in the scientific research. Moreover, the number of statistically significant *p-values* reported in scientific publications has increased over the years (Chavalarias et al., 2016) and there exists a tendency among researchers to look for that combination of data that provides a *p-value* smaller than 0.05 (Bruns and Ioannidis, 2016). However, the assessment of the *p-value* has some drawbacks which can lead to spurious scientific conclusions (Altman and Krzywinski, 2017; Amrhein et al., 2019; Bruns and Ioannidis, 2016; Halsey et al., 2015; Nuzzo, 2014). The data recorded from high-content, high-throughput studies, and the capacity of the computers to analyze thousands of numbers, has enabled us to enlighten the current uncertainty around the exploited *p-value*.

We report clear evidence about the well-known dependence of the *p-value* on the size of the data (Altman and Krzywinski, 2017; Krawczyk, 2015; Lin et al., 2013). While some approaches analyze the distribution of empirically estimated *p-values*, also known as *p-curve*, (Simonsohn et al., 2014), to the best of our knowledge, there are not approaches that focus on the size-dependence shown here to assess decision making. Due to the lack of new techniques to face the latter, we believe that our method will have a huge impact in the way scientists perform hypothesis testing.

The approximation of the function $p(n)$, through the use of a basic exponential function, lets us analyze the data more robustly utilizing $p(n)$ decay. With it, we provide a new perspective about hypothesis testing that prevents from treating the *p*-value as a dichotomous index and enables the study of data's variability. Using a simple mathematical formulation, a robust decision index θ is defined to enable good *praxis* in the

same context as statistical hypothesis testing. Indeed, the presented method is transferable to any field of study, same as the common null-hypothesis testing. Moreover, the presented approach used as a preliminary analysis, provides evidence about the existence (or not) of statistical significance. Even when large datasets are not available. Therefore, it supports the management of new data collection and can help researchers to reduce the cost of collecting experimental data.

The result of the pipeline (θ) relies on a new threshold called γ, which can only change in the most uncertain cases as shown in Materials and Methods. Compared to the classical *p-value* and α threshold, the parameter γ is mathematically constrained and θ is stable to its variations. Similarly, n_α is an unbiased indicator of the effect size, i.e. how different the samples are or how big is this difference (further details about robustness are given in Materials and Methods).

The computational cost of the proposed data diagnosis increases proportionally with the number of groups to compare. Therefore, the optimization of the code and its connection to either a GPU or cloud computing is recommended. Overall, we advocate for the implementation of our pipeline in user-friendly interfaces connected to either cloud-computing or GPU. The code provided within this manuscript is built into the free software Python, so that anyone with limited programming skills can include any change to obtain a customized tool.

ACKNOWLEDGEMENTS

This work was produced with the support of the Spanish Ministry of Economy and Competitiveness (TEC2015-73064-EXP, TEC2016–78052-RTC2017-6600-1, RTC-2017-6600-1), a 2017 Leonardo Grant for Researchers and Cultural Creators, BBVA Foundation, and grants from the US National Institutes of Health (UO1AG060903 and U54CA143868). We also want to acknowledge the support of NVIDIA Corporation with the donation of the Titan X (Pascal) GPU used for this research. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1746891. We thank Claire Jordan Brooks, Prof. Joachim Goedhart (University of Amsterdam), Laura Nicolás-Sáenz, and Pedro Macías-Gordaliza for fruitful discussions.

AUTHOR CONTRIBUTIONS

E.G.M. contributed to the conception and the implementation of the mathematical method, designed the data analysis experiments and wrote the manuscript with input from D.W. and A.M.B. A.S., H.J., J.M.P. and D.W. contributed with the experimental data. All authors contributed to the interpretation of the results. All authors reviewed the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

MATERIALS AND METHODS

Available in the following pages.

- DATA DETAILS
  - Drug analysis on phase contrast microscopy
- METHOD DETAILS
  - *p-value* as an exponential function of data size
  - Distance to the α-level of statistical significance
  - Mathematical formulation of the decision index
    - Restricting an optimal threshold
  - Data characterization in stable and uncertain cases.
  - Test of reliability

SUPPLEMENTAL INFORMATION available in the following pages.
CODE AVAILABILITY at https://github.com/BIIG-UC3M/pMoSS.

REFERENCES

Altman, N., and Krzywinski, M. (2017). P values and the search for significance. Nat. Methods *14*, 3–4.

Amrhein, V., Greenland, S., and McShane, B. (2019). Scientists rise up against statistical significance. Nature *567*, 305–307.

Bruns, S.B., and Ioannidis, J.P.A. (2016). P-curve and p-hacking in observational research. PLoS One *11*.

Chavalarias, D., Wallach, J.D., Li, A.H.T., and Ioannidis, J.P.A. (2016). Evolution of Reporting P Values in the Biomedical Literature, 1990-2015. JAMA *315*, 1141.

Cumming, G., Fidler, F., and Vaux, D.L. (2007). Error bars in experimental biology. J. Cell Biol. *177*, 7–11.

Halsey, L.G., Curran-Everett, D., Vowler, S.L., and Drummond, G.B. (2015). The fickle P value generates irreproducible results. Nat. Methods *12*, 179–185.

Khoury, P., Stokes, K., Gadkari, M., Makiya, M.A., Legrand, F., Hu, Z., Klion, A., and Franco, L.M. (2018). Glucocorticoid-induced eosinopenia in humans can be linked to early transcriptional events. Allergy *73*, 2076–2079.

Krawczyk, M. (2015). The Search for Significance: A Few Peculiarities in the Distribution of P Values in Experimental Psychology Literature. PLoS One *10*, e0127872.

Krzywinski, M., and Altman, N. (2013). Importance of being uncertain. Nat. Methods *10*, 809–810.

Leek, J., McShane, B.B., Gelman, A., Colquhoun, D., Nuijten, M.B., and Goodman, S.N. (2017). Five ways to fix statistics. Nature *551*, 557–559.

Lin, M., Lucas, H.C., and Shmueli, G. (2013). Research Commentary —Too Big to Fail: Large Samples and the p -Value Problem. Inf. Syst. Res. *24*, 906–917.

Mann, H.B., and Whitney, D.R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. Ann. Math. Stat. *18*, 50–60.

Nuzzo, R. (2014). Scientific method: Statistical errors. Nature *506*, 150–152.

Phillip, J.M., Wu, P.-H., Gilkes, D.M., Williams, W., McGovern, S., Daya, J., Chen, J., Aifuwa, I., Lee, J.S.H., Fan, R.,

et al. (2017). Biophysical and biomolecular determination of cellular age in humans. Nat. Biomed. Eng. *1*, 0093.

Simonsohn, U., Nelson, L.D., and Simmons, J.P. (2014). P-curve: A key to the file-drawer. J. Exp. Psychol. Gen. *143*, 534–547.

Smirnov, N. V (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. Bull. Math. Univ. Moscou *2*, 3–14.

Student (1908). The Probable Error of a Mean. Biometrika *6*, 1.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics Bull. *1*, 80.

Xu, Q.-S., and Liang, Y.-Z. (2001). Monte Carlo cross validation. Chemom. Intell. Lab. Syst. *56*, 1–11.

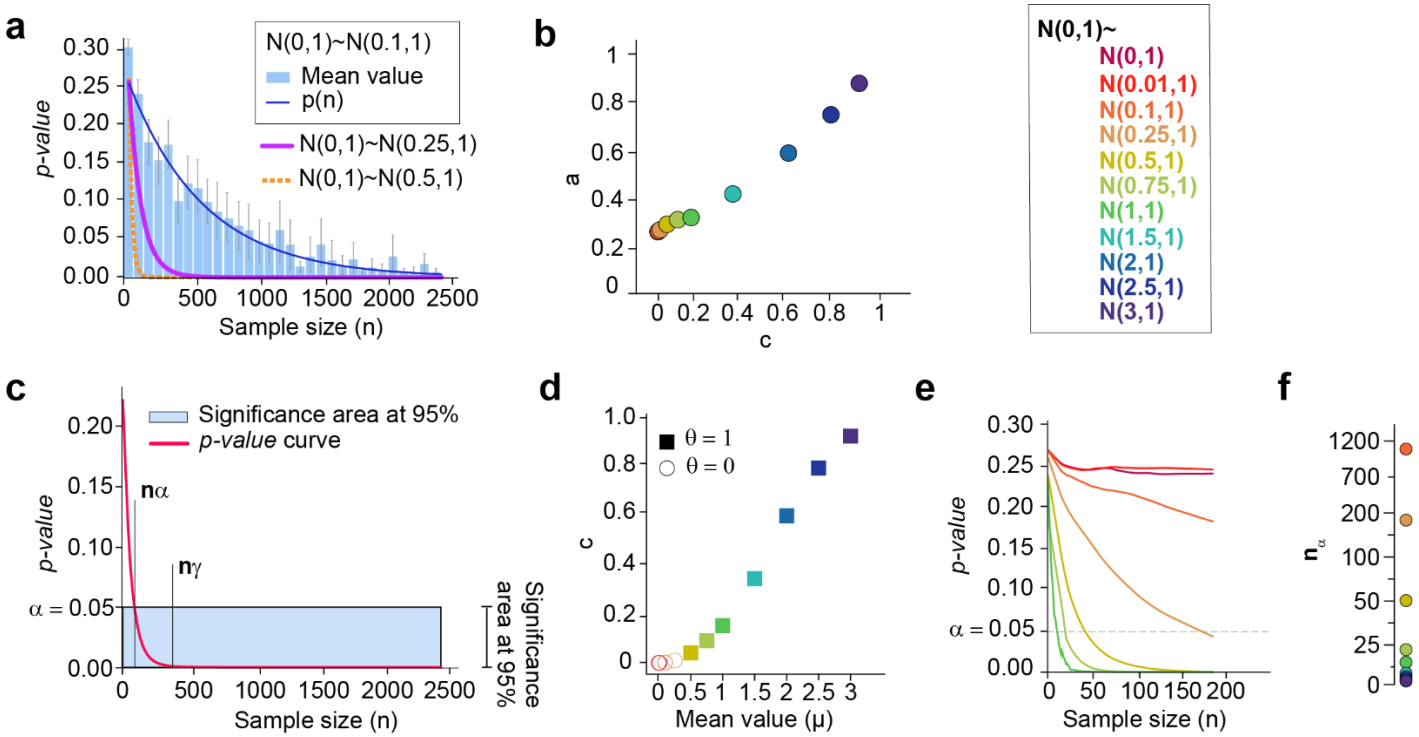

Fig. 1| Estimation of the *p-value* as a function of the size ($p(n)$) enables the correct discrimination between conditions. a) The *p-value* is a variable that depends on the sample size and can be modelled as an exponential function ($p(n) = ae^{-cn}$, Eq. 1). For each pair of normal distributions being compared, two subsets of size n are obtained by sampling from the corresponding normal distribution. Then, these datasets are compared using the Mann-Whitney statistical test and the *p-value* obtained is stored. The procedure is repeated many times for each size n. The blue bars with the standard error of the mean (SEM), show the distribution of all the *p-values* obtained at each size *n* when two normal distributions of mean 0 and 0.1, and standard deviation 1 are compared. The blue curve shows the corresponding exponential fit. The magenta and yellow curves represent the resulting $p(n)$ function when a normal distribution of mean 0 and standard deviation 1 is compared with a normal distribution of the same standard deviation and mean 0.25 and 0.5, respectively; **b**) The decay of $p(n)$ (parameters a and c of the exponential fit) increases with the mean value of the normal distribution being compared with $N(0,1)$. The larger the distances between the means of the distributions, the higher the decay of the exponential function (Table 1 in Materials and Methods). **c**) Comparison of $p(n)$ (red curve) and significance area at 95% (blue area). If the area under the red curve is smaller than the blue area, then there is a strong statistical significance. The parameter n_α measures the minimum data size needed to find statistical significance. The parameter n_γ measures the convergence of $p(n)$: $p(n = n_\gamma) \approx 0$. The binary decision index θ indicates whether the area under $p(n)$ from 0 to n_γ is larger than the area under the α-level (blue box) in the same range; **d**) The faster the decay of $p(n)$, the stronger the statistical significance of the tested null hypothesis. For $\gamma = 5e^{-06}$, $\theta_{\alpha,\gamma} = 1$ whenever the mean value of the normal distribution compared with $N(0,1)$ is larger than 0.5 (Table 1 in Materials and Methods). **e**) The empirical estimation of $p(n)$ with small datasets enables the detection of the most extreme cases: those in which the null hypothesis can be accepted, and those in which it clearly cannot; **f**) The minimum data size needed to obtain statistical significance (n_α) is inverse to the mean value of the normal distributions being compared.

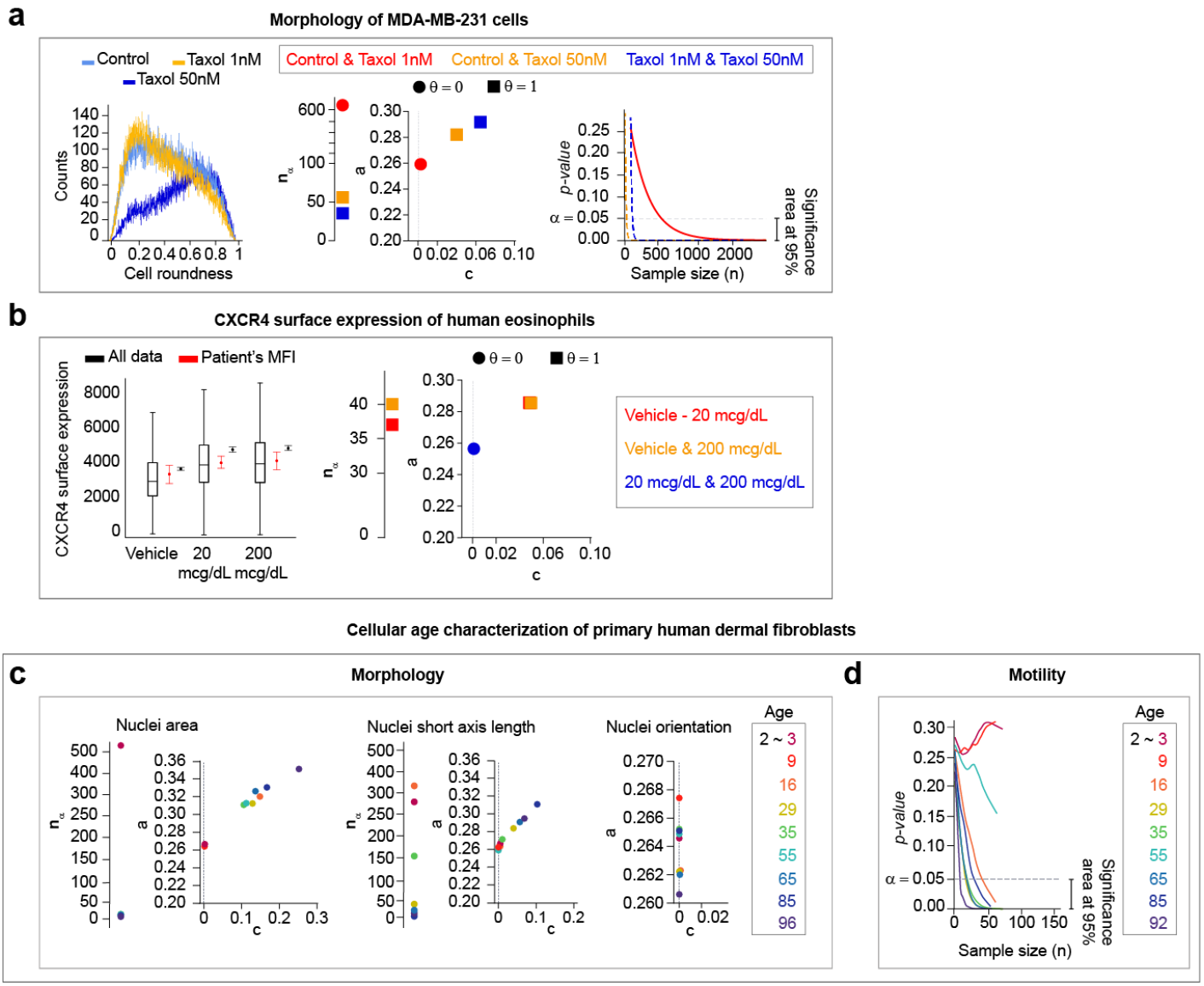

Fig. 2| The function $p(n)$ acts as a data descriptor and supports the experimental study of multiple conditions. a) Breast cancer cells (MDA-MB-231) were cultured in collagen and imaged under a microscope to determine if cells change shape when a chemotherapy drug (Taxol) is administered. Three different groups were compared: control (non-treated) cells, cells at 1 nM and at 50 nM Taxol. (Leftmost) The cell roundness distribution of control cells and cells treated at 1 nM Taxol have lower values than that of cells treated at 50 nM. (Right) The three groups were compared, the *p-values* were estimated and $p(n)$ was fitted for each pair of compared groups. When Taxol at 50 nM is evaluated (blue and yellow dashed curves), n_α is lower and the decay of $p(n)$ is higher (a and c parameters in Eq. 1), i.e. it decreases much faster than the one corresponding comparison of control and Taxol at 1 nM (orange curve). **b)** Flow cytometry data was recorded to determine the transcriptional changes induced by the in vivo exposure of human eosinophils to glucocorticoids. (Left) The entire dataset has a wider range of values and a smaller 95% confidence interval around the mean than the distribution obtained when the median fluorescence intensity (MFI) is calculated by each of the 6 subjects. (Right) There is an increase of the surface expression of CXCR4 when human eosinophils are exposed to 20 or 200 mcg/dL of Methylprednisolone. Namely, the minimum size n_α is low and the decision index $\theta = 1$ when any of those conditions are compared with the vehicle condition. Note that the decay parameters a and c are almost the same in those two cases, so the markers co-localize (Supplementary Material). The minimum size n_α when eosinophils are treated (blue circle) is not shown as it has infinite value. **c)** The morphology of 2 year-old human cells is compared with the morphology of 3, 9, 16, 29, 35, 45, 55, 65, 85 and 96 year-old human cells. For both, nuclei area and nuclei short axis measures, the minimum size n_α and the decay a

change proportionally with the age of the donor. The nuclei orientation does not characterize the age of the human donors for all the comparisons; the parameter c is null, and therefore, $p(n)$ is constant. **d)** The analysis of a small dataset is enough to determine that the total diffusivity can characterize the cellular aging in humans. The total diffusivity of 2, 3 and 9 year-old human cells are equivalent, while it differs when compared to cells from older human donors.

Materials and Methods

DATA DETAILS

Drug analysis on phase contrast microscopy

Phase contrast microscopy images of a human invasive ductal carcinoma (MDA-MB-231) cell line were acquired. The set-up used was composed by a Cascade 1K CCD camera (Roper Scientific), mounted on a Nikon TE2000 microscope with a 10X objective lens. Cells were embedded in 3D collagen type I matrix at 100.000 cells/mL. The time lapse videos were recorded every two minutes with a focus plane of at least 500 μm away from the bottom of the culture plates to diminish edge effects (He et al., 2017). Three different groups of cells were analyzed: control and treated with fresh media at 1 nM Taxol and 50 nM Taxol. Ten videos of 16.5 hours (500 frames of 809 μm x 810 μm with a resolution of 0.806 $\mu m/pixel$) each were analyzed per group.

All videos were automatically processed using a convolutional neural network (U-net (Ronneberger et al., 2015)) to get binary masks for the cell bodies and their protrusions. The resulting semantic segmentation corresponds uniquely to focused cells in the image. For each of these cells, their body and protrusions are segmented. See some examples of the resulting segmentation in Figure S2.

Using the segmentations, eight different morphological measurements were calculated: cell body size (CS), cell body perimeter (CP), cell body roundness (CR), cell with at least one protrusion (Pb), protrusion size (PS), protrusion perimeter (PP), protrusion length (PL) and protrusion diameter (PD) (Table S1). Further information about the distribution of each of the measurements is given in the Supplementary Material. Same as the analysis done for CR (Fig. 2a in the main manuscript), the differences between control and 1 mM and 50 nM Taxol were analyzed using each of the remaining variables (Figures S4-S6 and Table S3).

METHOD DETAILS

Here, we first provide the mathematical details behind our hypothesis that the *p-value* is a variable that critically depends on the size of the sample and that the *p-value* function can be approximated with an exponential function of the sample size n. Then, we define the method of how to work with the *p-value* as a function and to determine when a statement of statistical significance can be made ($\theta_{\alpha,\gamma}$, Eq. 9). Once the problem is described technically, it is possible to calculate the minimum size n_α at which the null hypothesis of the test is statistically significant (Eq. 11). This parameter n_α can be used to characterize the data. Finally, the reliability of our method is rigorously tested.

p-value as an exponential function of data size

Fig. 3 illustrates the idea that the *p-value* is a function that depends on the sample size n. There exists a

continuous inverse relation between *p-values* and n, i.e. p-values decrease when n increases, (Altman and Krzywinski, 2017; Krzywinski and Altman, 2013a, 2013b). This allows us to assume that *p-values* can be considered indeed, as a function of n, i.e. $p(n)$.

Either with Mann Whitney U test (Mann and Whitney, 1947) or with Student's *t*-test (Student, 1908), it can be proved that the obtained *p-value* converges to zero when the sample size is large and the distributions being assessed are not exactly the same, i.e., the *p-value* tends to zero when the sample size tends to infinity. A mathematical demonstration of this statement is available in the Supplementary Material.

Going a step further, we claim that the *p-values* can be indeed written directly as a function of n, $p(n)$, and that this function adjusts well to an exponential function. To show this, we first estimate the value that the *p-value* function has at each possible value of n. This can be done easily with the Monte Carlo cross validation method (MCCV) (Xu and Liang, 2001): at each iteration i of the procedure, $n = n_i$ is fixed, and two populations of size n_i are compared. This procedure is repeated many times in each given iteration i to cover the variability of the problem at $n = n_i$. At the end, we have as many sets of *p-values* as iterations i that are of the form:

$$\mathcal{P}_i = \{(n_i, p_i^j), j \in \mathbb{N}\},\ i \in \mathbb{N}\,. \qquad 1$$

Note that this procedure is similar to the upstrap (Crainiceanu, 2018) using an increasing fraction of the sample. The details about the procedure followed for the estimation of the *p-values* is explained in the Supplementary Material.

In Fig. 3, the procedure is applied using random populations from different normal distributions. We distinguish two different situations: either the obtained distributions are uniform, so the mean value of all the p_i values is constant for any i (Figs. 3a and 3b); or the distributions tend to decrease when the sample size n increases (Figs. 3c-f). In other words, $p(n)$ can be written as a continuous function. Hence, for each iteration i, each set of p_i values is averaged to obtain the empirical estimation of the function $p(n)$ at $n = n_i$ (red markers in Fig. 3). Then, a smooth curve is fitted to these values using locally weighted scatter plot smoothing (LOWESS) (Cleveland, 1979), which shows $p(n)$ has an exponential shape (Figs. 4a and 4b).

To prove that the estimated function $p(n)$ can be written as an exponential function, it is sufficient to verify that the quotient between its first derivative $\frac{\partial p(n)}{\partial n}$ and the function $p(n)$ is itself a constant, i.e.

$$\frac{\frac{\partial p(n)}{\partial n}}{p(n)} = c \leftrightarrow p(n) = a \cdot e^{cn}\ where\ a, c \in \mathbb{R}\,. \qquad 2$$

Collecting the values $p(n)$ of the LOWESS fit, the quotient $\frac{\frac{\partial p(n)}{\partial n}}{p(n)}$ is calculated (Figs. 4c and 4d). Most of the quotients verify the condition in Eq. 2. In Fig 4c, we show cases in which it is more challenging to decide

whether there exists a statistical difference, as for instance, when $N(0,1)$ and $N(0.1,1)$ are compared. When $p(n)$ becomes very small, the quotient $\frac{\frac{\partial p(n)}{\partial n}}{p(n)}$ has more outliers, especially when the sample size n is small. This can be observed when comparing $N(0,1)$ with $N(0.75,1)$, $N(1,1)$, $N(2,1)$ and $N(3,1)$. (Fig. 4d). These are extreme cases in which there exist clear differences between populations and therefore, *p-values* are close to zero most of the time.

As we have proved above that the estimated function $p(n)$ can be written as an exponential function, an exponential curve is fitted to all the values p_i calculated with MCCV (Figs. 4a and 4b). Both LOWESS and exponential curves are very close to each other, even if the former was fitted using the mean values of each group p_i and the latter with all of them. An exponential fit is more suitable in this case as it is calculated with all the values obtained through MCCV, and only outputs positive values by definition. A LOWESS approximation can occasionally lead to biased negative values, such as when $N(0,1)$ and $N(0.75,1)$ are compared while the *p-values* are positively defined. Note that as $p(n) \rightarrow 0$ when $n \rightarrow \infty$, $c < 0$ necessarily in Eq. 2. Therefore, we assume from now on that $p(n)$ can be given as an exponential function of the form

$$p(n) \approx a \cdot e^{-cn} \; where \; a, c \in \mathbb{R}^{+}. \qquad 3$$

Here the parameters a and c control the amplitude and the decay of the function $p(n)$, respectively. If $c = 0$, then the value of $p(n)$ would be uniform in a: $p(n) = a$. As *p-values* are computed probabilities and the global maximum of $p(n)$ is a, a belongs to the [0, 1] interval.

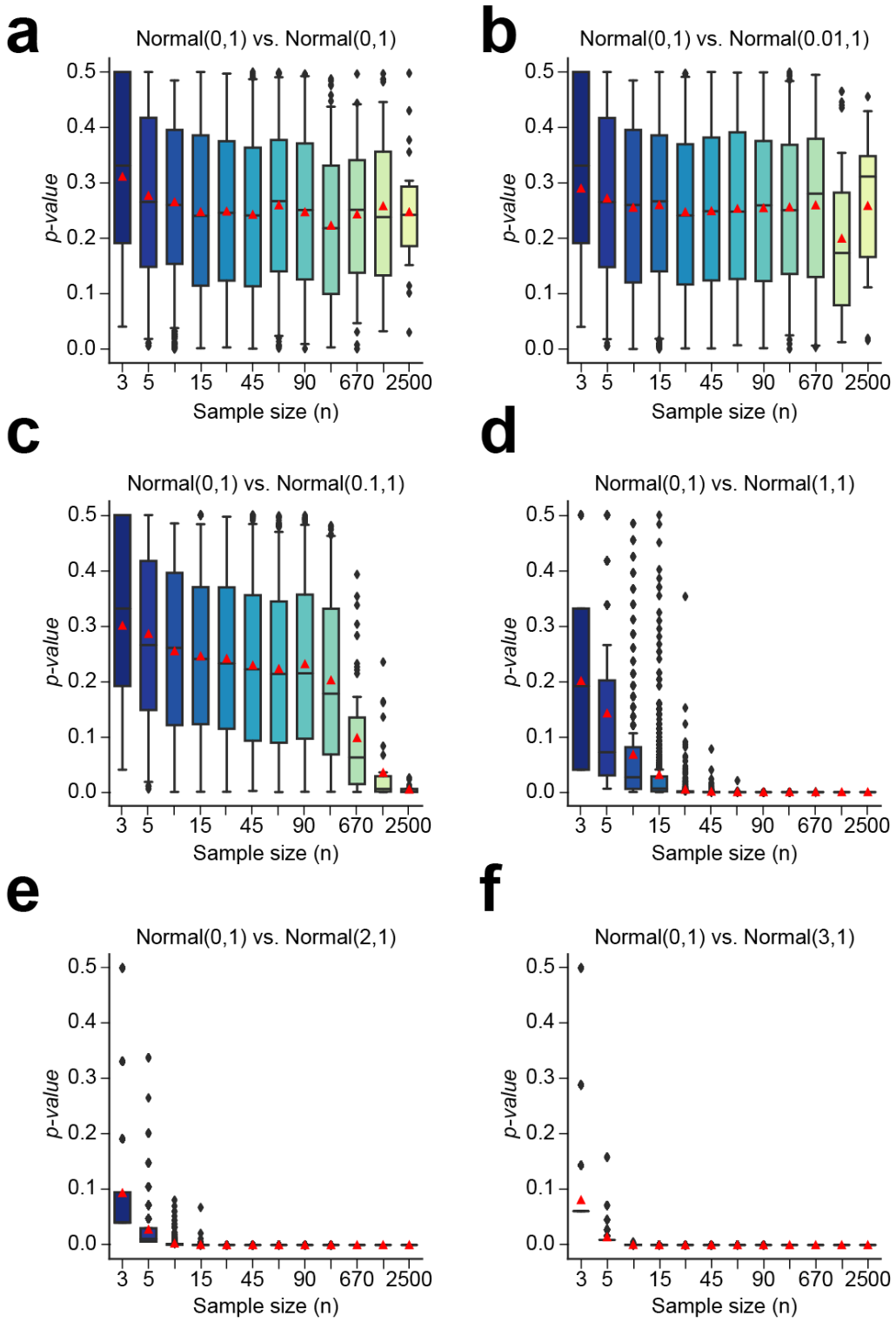

Figure 3. Distribution of the *p-values* obtained when two normal distributions are compared. For each fixed value of the sample size (3, 5, 10, 15, 30, 45, 60, 90, 200, 670, 1750 and 2499 points), two normal distributions of that size are simulated and compared via the Mann-Whitney statistical test. This procedure is repeated multiple times. A normal distribution with a mean of 0 and a standard deviation of 1 is compared with a normal distribution of mean: **(a)** 0, **(b)** 0.01, **(c)** 0.1, **(d)** 1, **(e)** 2, and **(f)** 3 and a standard deviation of 1. When both normal distributions are almost the same, **(a)** and **(b)**, the *p-value* follows a uniform distribution. Though, as long as both normal distributions get farther to each other, the distribution of *p-values* become closer to a normal distribution with a faster decay.

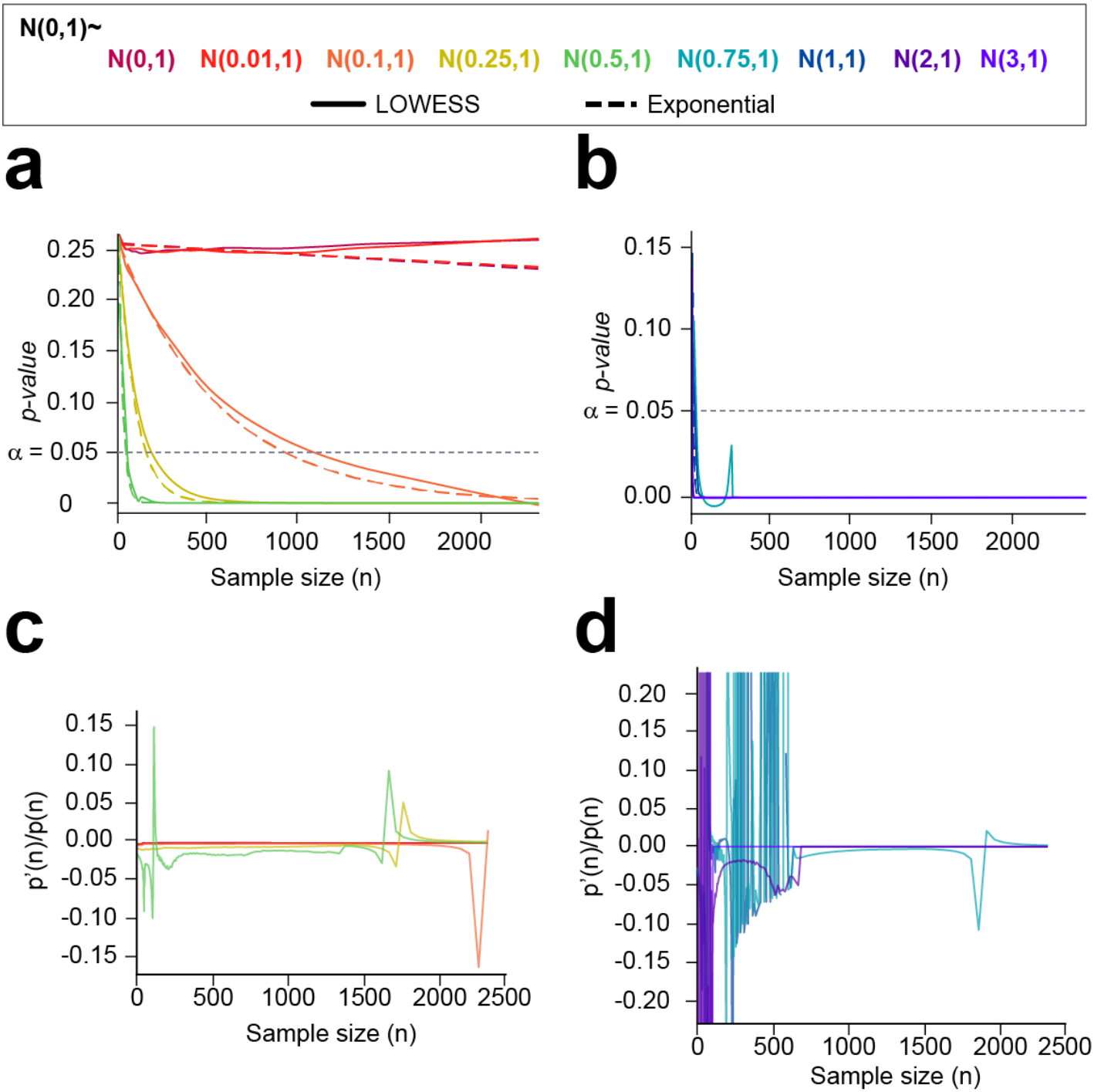

Figure 4. A normal distribution with a mean of 0 and a deviation of 1 is compared with a normal distribution of mean (0, 0.01, 0.1, 0.25, 0.5, 0.75, 1, 2 and 3). Multiple *p-values* are calculated for a sample size between two and 2500 (Fig. 3). **(a)** and **(b)** Locally weighted scatter plot smoothing (LOWESS) fit to the mean *p-values* (red markers in Fig. 3) computed for each value of the sample size n. Likewise, an exponential function is fitted to all the simulated *p-values*. **(c)** and **(d)** Quotient between each LOWESS curve and its differential. **(c)** Comparison of $N(0,1)$, with $N(0,1)$, $N(0.01,1)$, $N(0.1,1)$, $N(0.25,1)$ and $N(0.5,1)$. **(d)** $N(0,1)$ is compared with $N(0.75,1)$, $N(1,1)$, $N(2,1)$ and $N(3,1)$. Constant quotients and accurate exponential fits show empirically that $p(n)$ has an exponential nature.

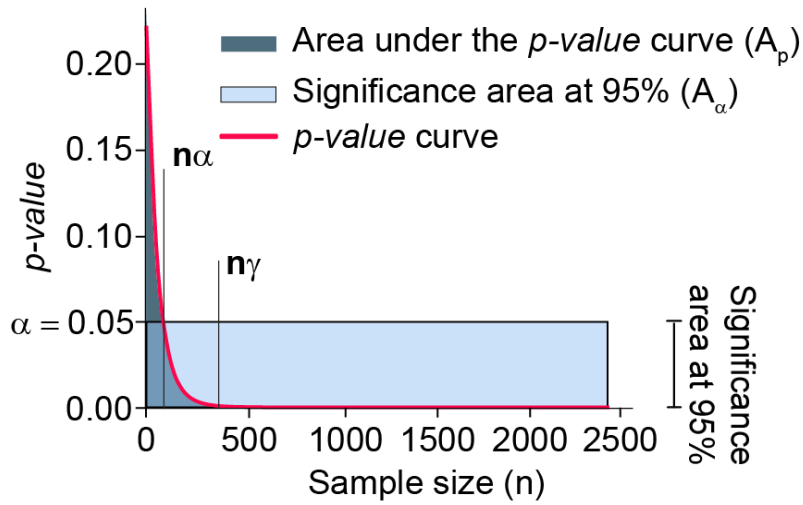

Figure 5. Comparison of a 95% of statistical significance ($\alpha = 0.05$) and an *n*-dependent *p-value* curve. The parameter n_α represents the minimum sample size to detect statistically significant differences among compared groups. The parameter n_γ represents the convergence point of the *p-value* curve. When the *p-value* curve expresses statistically significant differences, the area under the red curve ($A_{p(n)}$) is smaller than the area under the constant function $\alpha = 0.05$ ($A_{\alpha=0.05}$) when it is evaluated between 0 and n_γ.

Distance to the α-level of statistical significance

The ideal case of a true $(1 - \alpha)$ statistical significance would lead to the rejection of the null hypothesis independently of data size, i.e., *p-values* would always be lower than α. Hence, we claim that whenever there exist real statistically significant differences between two samples, $p(n)$ reaches α rapidly. So, the values of $p(n)$ are mostly distributed in a range smaller than α. Therefore, we compare all the values of the curve $p(n)$ with α. In the discrete case, we would evaluate $\alpha - p(n = n_i)$ for each index i and sum all the results: if the sum is positive, then $p(n)$ is smaller than α most of the time. In the continuous case, this sum is obtained by integrating the difference

$$\delta_\alpha(n) = \int \big(\alpha - p(n)\big) dn = A_{\alpha(n)} - A_{p(n)}, \qquad 4$$

where A_α is the area under the constant function α and $A_{p(n)}$ is the area under the estimated *p-values*' curve, $p(n)$ (Fig. 5). A positive $\delta(n)$ implies that A_α is larger than $A_{p(n)}$, i.e. most of the values in $p(n)$ are below the significance threshold α; a negative $\delta(n)$ implies the opposite.

As shown in the next paragraphs, Eq. 4 aims to quantify and evaluate the distribution of *p-values* (i.e., the distribution of $\{(n, p(n)), n \in N\}$) taking into account two aspects, whether (1) most of the *p-values* are smaller than α and (2) the decay of $p(n)$ is large enough.

Mathematical formulation of the decision index

By means of the exponential expression of $p(n)$ given in Eq. 3, the measure $\delta_\alpha(n)$ (Eq. 4) can be rewritten as follows

$$\delta_\alpha(n) = \alpha n - \frac{a}{c}(1 - e^{-cn}). \qquad 5$$

Due to the limits of a and c, $\delta_\alpha(n)$ is still well-defined. However, in the limit of n, $\delta_\alpha(n)$ will always be positive and it tends to infinity:

$$\lim_{n\to\infty} \delta_\alpha(n) \approx \lim_{n\to\infty} \left(\alpha n - \frac{a}{c}\right) \to \infty. \qquad 6$$

Also, from a practical perspective, the area of interest to evaluate the decay of $p(n)$ is that enclosed between zero and its convergence point n: $\left|\frac{\partial p(n)}{\partial n}\right| \approx 0$. Namely, a relevant sub-sample size n can be computed as

$$n_\gamma = argmin_{\,n} \left\{ \left|\frac{\partial p(n)}{\partial n}\right| < \gamma \right\}, \qquad 7$$

where γ is the threshold chosen to determine the convergence point (Fig. 5). Finally, $\delta_{\alpha,\gamma}$ is now formally defined as

$$\delta_{\alpha,\gamma} = A_{\alpha,\gamma} - A_{p(n=n_\gamma)} = \alpha n_\gamma - \frac{a}{c}(1 - e^{-cn_\gamma}). \qquad 8$$

As claimed at the end of the last section, the computation of $\delta_{\alpha,\gamma}$ enables the identification of a rapid convergence to zero at small values of n induced by the high slope of $p(n)$, which is indicative of the existence of true statistical significant differences.

The decision index we propose, $\theta_{\alpha,\gamma}$, is defined as

$$\theta_{\alpha,\gamma} = \begin{cases} 1, & \delta_{\alpha,y} \geq 0 \\ 0, & otherwise \end{cases}, \qquad 9$$

where $\delta_{\alpha,\gamma}$ follows Eq. 8.

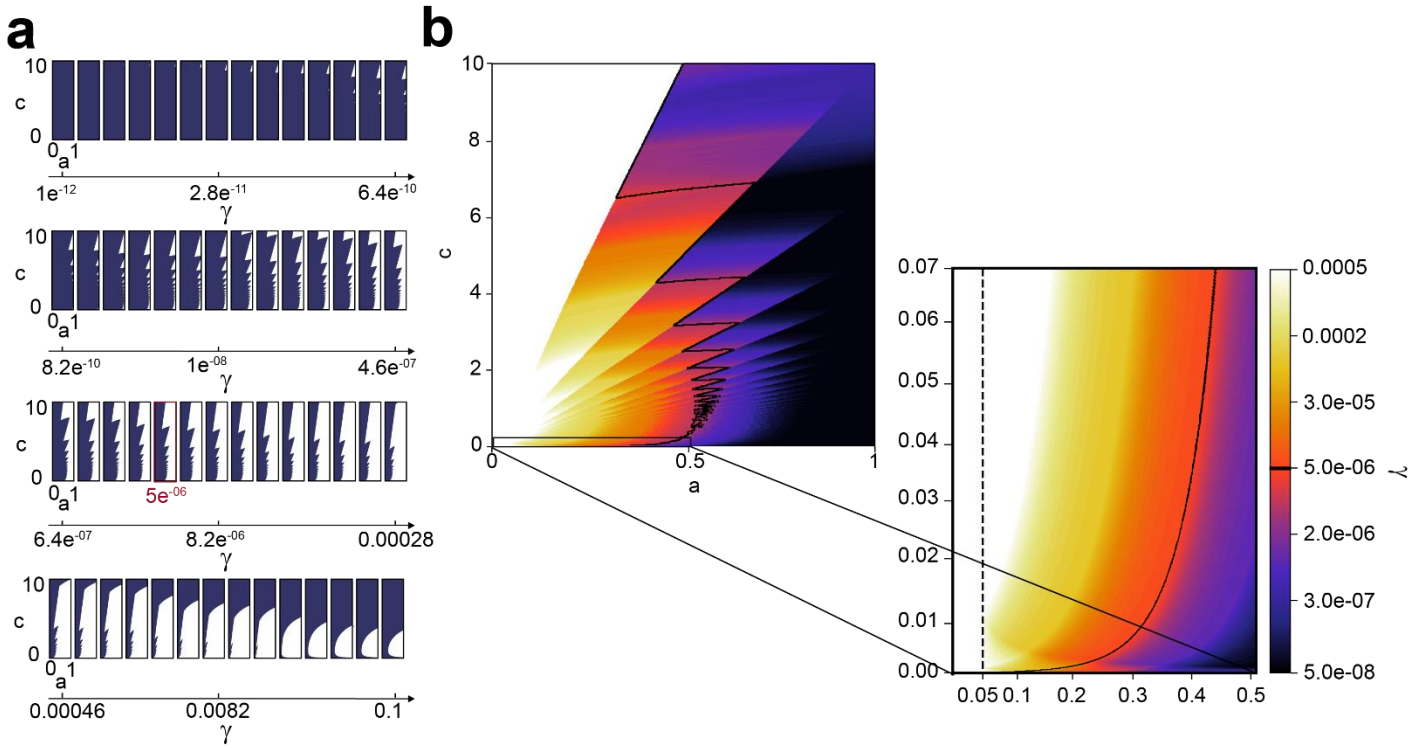

Figure 6. Decision index $\theta_{\alpha=0.05,\gamma}$ for different values of parameters a and c in the function ae^{-cn} and threshold γ: **(a)** Each of the subplots is drawn for a specific value of γ, being the dark area the cases for which there exists a 95% of statistical significance ($\theta_{\alpha=0.05,\gamma} = 1$), and white area the rest of the cases $\theta_{\alpha=0.05,\gamma} = 0$; **(b)** Colors in the image correspond to the values of γ for which $\delta_{\alpha=0.05,\gamma} = 0$. The black frontier shows $\delta_{\alpha=0.05,\gamma=5e^{-06}} = 0$ (red box in **(a)**). All the values of a and c for which $\theta_{\alpha=0.05,\gamma=5e^{-06}} = 1$ (95% of statistical significance) lie on the left side of this limit and, the rest, on the right. The plots shown in **(a)** show the influence of the parameter γ in a wide range of values, while the plots shown in **(b)** are limited to the range of values we find in this posterior experiment. The vertical dashed line indicates the cases $a = 0.05$ which are always statistically significant.

Restricting an optimal threshold

The proposed approach depends on two thresholds: (1) significance threshold α and (2) the convergence threshold γ. The former measures the level of statistical significance, while the latter controls decisions. Therefore, the only critical threshold to discuss in this work is γ.

The rules to follow for the selection of the threshold γ are:

- The parameter a is the maximum value that $p(n)$ can take. Therefore, if a is smaller than α, then $\theta_{\alpha,\gamma} = 1$ for any γ given.
- As $\delta_{\alpha,\gamma}(n)$ tends to infinity with n, the smaller the value of γ is set, the larger n_γ will be and the chances of $\theta_{\alpha,\gamma} = 1$ will also increase.
- The values of γ should be small: α is considered a significant number and $p(n)$ values are constantly compared with it. It seems reasonable to compare the slope of $p(n)$ at the convergence point with a value smaller than α, which is usually smaller than 0.1.

Eq. 7 implies

$$\left|\frac{\partial p(n)}{\partial n}\right| = |-cp(n)| < \gamma \leftrightarrow p(n) < \frac{\gamma}{|c|}. \qquad 10$$

So, if γ is chosen such that $\frac{\gamma}{|c|}$ is greater than α, it would vanish the assumption that $p(n)$ has arrived to a convergence point equivalent to zero. Therefore, our claim is that $\frac{\gamma}{|c|} < \alpha$ with at least $\gamma < 0.1$.

Data characterization in stable and uncertain cases.

The threshold γ controls severe decisions. Namely, the lower this value is set, the less strict the decision will be. In Fig. 6a, we show the dynamics of $\theta_{\alpha=0.05,\gamma}$ when γ changes: the dark area ($\theta_{\alpha=0.05,\gamma} = 1$) increases inversely to γ, showing that the chances for which the null hypothesis is rejected increase as well. Moreover, the limit between dark and light ($\theta_{\alpha=0.05,\gamma} = 0$) areas is precisely the curve $\delta_{\alpha,\gamma} = 0$. The value of γ determines this curve and therefore, the conditions for which $\theta_{\alpha=0.05,\gamma} = 1$ (dark area) and $\theta_{\alpha=0.05,\gamma} = 0$ (light area). In Fig. 6b, we illustrate the condition $\delta_{\alpha,\gamma} = 0$ when $\alpha = 0.05$, as a function of a, c and γ. The case $\gamma = 5e^{-06}$ is underlined in black.

There exist some points (a, c) for which the rejection of the null hypothesis is independent of γ. A clear example is the case in which $a \geq \alpha$ and $c \approx 0$. These cases represent the situation in which the null hypothesis cannot be rejected with a statistical significance of level α. For instance, when $N(0,1)$ is compared with $N(0,1)$ or $N(0.01,1)$ (Fig. 1b in the main manuscript, Figs. 3a-b). Likewise, if $a \leq \alpha$ or c is large enough, the null hypothesis is always rejected with a statistical significance of level α. For instance when $N(0,1)$ is compared with $N(2,1)$ or $N(3,1)$ (Fig. 1b in the main manuscript, Figs. 3e-f).

The proposed methodology let us also classify each case by its level of uncertainty. The coefficients to fit an exponential curve are precisely coordinate points in any of the plots in Fig. 6. Therefore, once an exponential curve is fitted and parameters a and c are known, it is possible to know in which position of the graph is the case of study: clear cases will always be close to the left or to the right side of the graphs in Fig. 6, while most unstable or unclear cases will be placed in the middle. Therefore, with this method, it is possible to determine when there are statistically significant differences, and when these differences are not sufficiently clear and it might be necessary to perform a deeper study.

An intuitive interpretation of statistical significant differences between two groups (the classical threshold *p-value* $< \alpha$) is that their mean confidence intervals do not overlap. These intervals decrease when the size of the data increases (Krzywinski and Altman, 2013c). Therefore, this section is devoted to study how large two populations must be in order to obtain non-overlapping intervals. Interestingly, the estimation of the function $p(n)$ allows us to determine the specific minimum value of n, n_α, for which $p(n)$ is lower than the significance level α (Fig. 5). This value is the solution to the equation

$$\alpha = ae^{-cn_\alpha}. \quad 11$$

As computed, n_α represents the minimum sample size needed to obtain a statistically significant *p-value*, in case it exists. In other words, reproducing an experiment with n_α samples assures the rejection of the null hypothesis. The estimated n_α allows to assess the strength of the evidence against the null hypothesis. If n_α is small, the strength of the statistical difference is very clear and two populations are distinguishable.

The parameters a and c in Eq. 11 are obtained empirically through MCCV so they can introduce some bias in the calculation of n_α. Hence, a better estimator of n_α, $\hat{n}_\alpha$, can be computed using the *p-values* obtained directly from the data and their variance

$$\hat{n}_\alpha = argmin_{\,n_i}\{\,(\overline{p_i} - \sigma_{\overline{p_i}}) < \alpha\}, \quad 12$$

where $\overline{p_i}$ represents the mean of the set of values p_i (MCCV) and $\sigma_{\overline{p_i}}$, the mean standard error (SEM), which is included to correct for the variability of the estimated *p-values*. The estimator $\hat{n}_\alpha$ is limited to those cases in which the data is large enough: if the size of the data is smaller than n_α, then $\hat{n}_\alpha$cannot be computed (Fig. 1e-f, Fig. 2d in the main manuscript). As $\hat{n}_\alpha$ is more restrictive than n_α, its value will always be slightly larger (Table 1).

Test of reliability

Unlike many computational methods, the analysis of statistical significance of the differences between two groups cannot be evaluated by means of Ground Truth data, simulations or human-made annotations. Nonetheless, it is possible to determine the robustness on the reproducibility of the results. Namely, whether the statistical significance is maintained when the experiment is repeated. To do so, we test our method using simulated normal distributions.

Any data diagnosis carried out with the proposed method depends on the value γ chosen and the limitations posed by its computational intensive nature. As done at the beginning of this work, we compare the normal distribution $N(0,1)$ with $N(0.01,1)$, $N(0.1,1)$, $N(0.25,1)$, $N(0.5,1)$, $N(0.75,1)$, $N(1,1)$, $N(2,1)$ and $N(3,1)$. We should obtain $\theta_{\alpha,\gamma} = 1$ when comparing the most similar distributions such as $N(0,1)$ and $N(0.01,1)$. In contrast, we should get $\theta_{\alpha,\gamma} = 0$ when comparing the most different distributions, such as $N(0,1)$ and $N(2,1)$.

To evaluate the effect of γ, $p(n)$ is simulated for all pairs of normal distributions and it is compared with a significance level of $\alpha = 0.05$ using different values of γ (Table S7 in the Supplementary Material). The lower the convergence criteria γ is, the less restrictive the diagnosis is (Fig. 6). Using the simulated data, the range of $\theta_{\alpha=0.05,\gamma}$ values obtained let us recommend a value for this parameter. When $N(0,1)$ and $N(0.1,1)$ are compared with $\gamma = 2.5e^{-06}$, the decision index $\theta_{\alpha=0.05,\gamma=2.5e^{-06}} = 1$ indicates that there exist

statistically significant differences among both distributions, which is the opposite of what we expected. If the value of parameter γ increases, the statistical significance is rejected in those cases in which there is a larger uncertainty. For instance, when $N(0,1)$ and $N(0.25,1)$ are compared with $\gamma = 5e^{-05}$, $\theta_{\alpha=0.05,\gamma=5e^{-05}} = 0$. However, the latter is not straightforward for two reasons: $\delta_{\alpha=0.05,\gamma=5e^{-05}} = -5.84$ (small difference) and $\hat{n}_{\alpha} = 186$ (few samples to observe statistically significant differences). Therefore, we strongly recommend the use of $\gamma = 5e^{-06}$.

Comparison	a	c	$\hat{n}_{\alpha}$	n_{γ}	$\theta_{\alpha=0.05,\gamma=5e^{-06}}$
$N(0,1)\sim N(0,1)$	0.256	0.000	∞	39599	0
$N(0,1)\sim N(0.01,1)$	0.255	0.000	∞	44237	0
$N(0,1)\sim N(0.1,1)$	0.257	0.002	1192	988	0
$N(0,1)\sim N(0.25,1)$	0.263	0.010	185	165	0
$N(0,1)\sim N(0.5,1)$	0.286	0.042	47	41	1
$N(0,1)\sim N(0.75,1)$	0.304	0.091	20	19	1
$N(0,1)\sim N(1,1)$	0.313	0.152	13	12	1
$N(0,1)\sim N(1.5,1)$	0.411	0.344	7	6	1
$N(0,1)\sim N(2,1)$	0.579	0.599	5	4	1
$N(0,1)\sim N(2.5,1)$	0.738	0.794	4	3	1
$N(0,1)\sim N(3,1)$	0.867	0.924	4	3	1

Table 1. Parameters of the function p(n) after the exponential fit with $\alpha = 0.05$ and $\gamma = 5e^{-06}$, for the comparison of a normal distribution with mean value 0 and standard deviation 1, and normal distributions of mean value 0, 0.01, 0.1, 0.25, 0.5, 0.75, 1, 1.5, 2, 2.5 and 3.

To test the generality of this results, the same procedure was repeated several times by changing the samples of the normal distributions being compared. Hence, it is possible to provide a probability of how often the resulting $\theta_{\alpha,\gamma}$ would be the same as the one stated in Table 1. Additionally, the presented method has its limitations in the computational time needed to perform MCCV iterations: the more iterations we compute the longer the process will take. Moreover, the accuracy of any estimated $p(n)$ depends on the sample size $n = n_i$ and *p-values*, p_i, that the program can evaluate. Therefore, we also tested the results of the method when the number of iterations i and j in MCCV is reduced. Overall, the probability of obtaining exactly the same result under any change of the previous conditions was calculated (Table S8 in the Supplementary Material). The closer this probability gets to 100, the more robust and general the result will be. We can confirm that the results are most of the time the same as the ones given in Table 1 when $\gamma = 5e^{-06}$. The only critical case is the comparison $N(0,1)$ - $N(0.5,1)$ when few n_i points are used to estimate $p(n)$.

The last procedure was repeated using the real data from the first experiment (study of the effect of Taxol

in the cell body and protrusions morphology) (Tables S9 and S10 in the Supplementary Material). Even with more complex and noisier data, the results obtained show that the method is stable and robust. All technical details about these computations are given in the Supplementary Material.

REFERENCES

Altman, N., and Krzywinski, M. (2017). P values and the search for significance. Nat. Methods *14*, 3–4.

Cleveland, W.S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. J. Am. Stat. Assoc. *74*, 829–836.

Crainiceanu, C.M. (2018). The upstrap.

He, L., Sneider, A., Chen, W., Karl, M., Prasath, V., Wu, P.-H., Mattson, G., and Wirtz, D. (2017). Mammalian Cell Division in 3D Matrices via Quantitative Confocal Reflection Microscopy. J. Vis. Exp. *56364*.

Krzywinski, M., and Altman, N. (2013a). Error bars. Nat. Methods *10*, 921–922.

Krzywinski, M., and Altman, N. (2013b). Significance, P values and t-tests. Nat. Methods *10*, 1041–1042.

Krzywinski, M., and Altman, N. (2013c). Importance of being uncertain. Nat. Methods *10*, 809–810.

Mann, H.B., and Whitney, D.R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. Ann. Math. Stat. *18*, 50–60.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, ed. (Springer International Publishing), pp. 234–241.

Student (1908). The Probable Error of a Mean. Biometrika *6*, 1.

Xu, Q.-S., and Liang, Y.-Z. (2001). Monte Carlo cross validation. Chemom. Intell. Lab. Syst. *56*, 1–11.

Confronting *p-hacking*: addressing *p-value* dependence on sample size.

Supplementary Information

E. Gómez-de-Mariscal, A. Sneider, H. Jayatilaka, J. M. Phillip, D. Wirtz and A. Muñoz-Barrutia

Corresponding author: Arrate Muñoz-Barrutia
E-mail: mamunozb@ing.uc3m.es
Code availability at https://github.com/BIIG-UC3M/pMoSS

Contents

1 Technical details **1**
- A *p-values* tend to zero for large sample sizes . . . 1
- B *p-values* as a function of the sample size . . . 2
- C Estimation of *p-values* with Monte Carlo cross validation method . . . 2
- D Assessment of minimum data size needed for statistical significance (n_α) . . . 3

2 Experimental data **4**
- A Experiment 1: Drug analysis on phase contrast microscopy data . . . 4
- B Experiment 2: Drug analysis on flow cytometry data . . . 5
- C Experiment 3: Cellular age characterization by means of biomolecular and biophysical properties . . . 6

3 Test of robustness **8**
- A Test of robustness on theoretical data . . . 8
- B Test of robustness on real data . . . 8
  - B.1 The *p-value* can be estimated by an exponential function. . . . 8
  - B.2 Robustness of the convergence threshold and required computational load . . . 9

1. Technical details

The main motivation of the study is that the *p-value* is no longer useful when working with large datasets as its value tends to zero. In the next section, we demonstrate for the particular cases of the Mann-Whitney U test, (1), and Student's t-test, (2), that indeed, the *p-value* will always tend to zero even when the null hypothesis is almost true and should not be rejected.

A. *p-values* tend to zero for large sample sizes. The statistic $\mathcal{U}$ of the Mann-Whitney U test, (1), is defined as $\min\{U_1, U_2\}$, where U_i follows the Eq. (1), being n_i the size of the dataset i and R_i its rank sum.

$$U_i = n_1 n_2 + \frac{n_i(n_i+1)}{2} - R_i, \quad i \in \{1,2\}. \qquad [1]$$

When n_i are large enough, $\mathcal{U}$ follows a normal distribution, (1), with mean and standard deviation values, $\mu_\mathcal{U}$ and $\sigma_\mathcal{U}$ respectively, described by Eq. (2).

$$\mu_\mathcal{U} = \frac{n_1 n_2}{2}, \quad \sigma^2_\mathcal{U} = \frac{n_1 n_2 (n_1+n_2+1)}{12}. \qquad [2]$$

Therefore, the main procedure to estimate the *p-value* consists in analyzing the typified value of $\mathcal{U}$, z, defined by

$$z = \frac{\mathcal{U} - \mu_\mathcal{U}}{\sigma_\mathcal{U}}. \qquad [3]$$

Replacing the values of U_i, $\mu_\mathcal{U}$ and $\sigma^2_\mathcal{U}$ in Eq. (3), we obtain

$$z = \sqrt{12}\left(\frac{n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 - \frac{n_1 n_2}{2}}{\sqrt{n_1 n_2 (n_1+n_2+1)}}\right). \qquad [4]$$

Note that U_1 and U_2 can be indifferently chosen to be the minimum value for the Mann-Whitney statistic $\mathcal{U}$. Hence, for simplicity $\mathcal{U} = U_1$ is assumed.

In the worst case scenario, when both datasets are identical and therefore the null hypothesis should be true, $R_1 = R_2 = R$. Also, as n_i are assumed to be large enough, we can study the case $n_1 = n_2 = n$. Moreover, due to the hypothesized large sample size, R_i could be upper limited as

$$R \leq \sum_{i=1}^{n} i = \frac{n(n+1)}{2} \qquad \Longrightarrow \qquad z \geq \sqrt{12}\left(\frac{\frac{n^2}{2}}{\sqrt{n^2(2n+1)}}\right) = \frac{n\sqrt{3}}{\sqrt{(2n+1)}}. \qquad [5]$$

Finally, the value of z in the limit, when n tends to infinity, is also infinity

$$\lim_{n\to\infty} z \geq \lim_{n\to\infty} \frac{n\sqrt{3}}{\sqrt{(2n+1)}} \longrightarrow \infty \qquad \Longrightarrow \qquad \lim_{n\to\infty} z \longrightarrow \infty. \qquad [6]$$

Therefore, *p-value* tends to zero. That is to say, even when we assume that both datasets are equal, the result would be to regret the null hypothesis. Likewise, Student's t-test (2) fails by means of large samples. The statistic $\mathbf{t}$ is defined as follows

$$\mathbf{t} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{S^2}{n_1} + \frac{S^2}{n_1}}}, \qquad [7]$$

where μ_i and n_i correspond to the mean and sample size of the dataset $i \in \{1, 2\}$. Once again, assuming that both n_i are large enough, $n_i = n$ is accepted; $\mathbf{t}$ is directly compared with the Student's t distribution and in the limit of n, t tends to infinity (as long as both mean values are not exactly the same). Thus, *p-value* tends to zero and the null hypothesis is rejected.

B. *p-values* as a function of the sample size. Proofs in Section A let us concluding that the *p-values* depend on the size of the data being evaluated. While this is not a breakthrough, it is one of the pillars in this study. The fact that the *p-value* varies with n, allows us to assume that they can be considered indeed, as a function of n. In the case of Student's t-test, it is straightforward that the t parameter is n-dependent (Eq. 7). Note that mean and standard deviation values are similar for any n.

In the case of the parameter z, Eq. 3, it can be slightly more complicated to prove the same statement. However, it is easy to see that $\mathcal{U}$, $\mu_{\mathcal{U}}$, and $\sigma_{\mathcal{U}}$ depend on n, and that z will always increase with respect to n (i.e., *p-value* decreases). Therefore, we can assume that the estimated *p-values* can be written as a function of n.

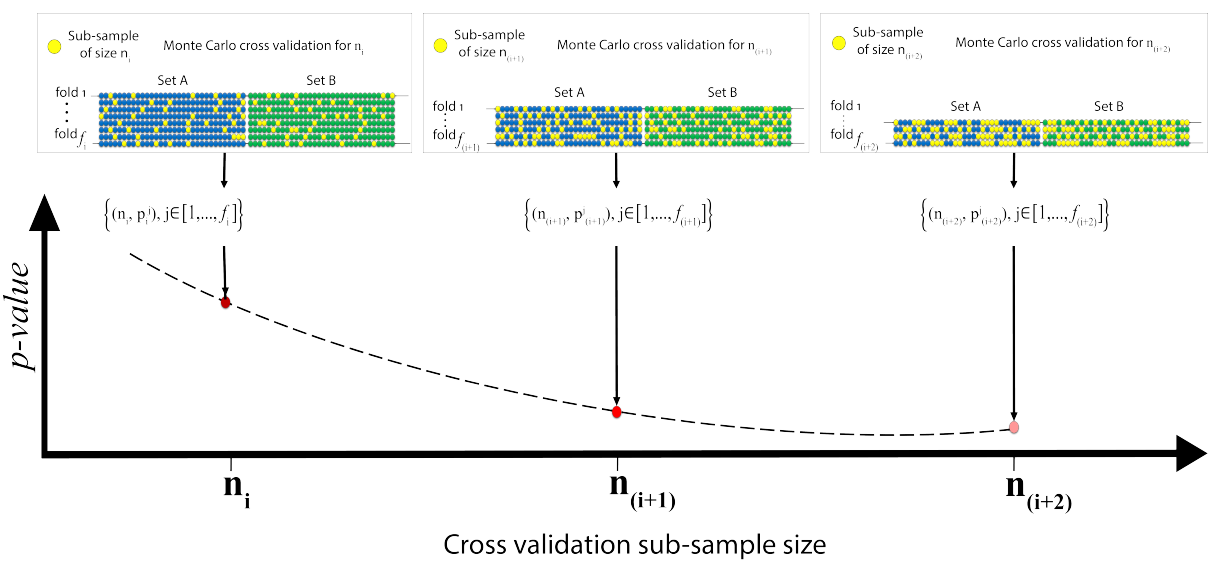

Fig. S1. Illustration of the work flow used for the estimation of *p-values* as a function ($p(n)$) of the sample size (n). For each possible value of n (n_i), Monte Carlo cross validation (MCCV) is performed f_i times. For each fold in the cross validation, two random sub-samples of size n_i are chosen from samples A and B (yellow spheres). Then, a statistical test is applied to obtain a *p-value* (p_i^j). The procedure is repeated f_i times; f_i depends on n_i as both samples A and B have to be covered. Thereby, f_i decreases ($f_i > f_{i+1} > f_{i+2}$) as long as n_i gets larger. This procedure is repeated for n_0, ..., n_{i+1},n_{i+2}, ... until the desired data size ($n_{i\to\infty}$) is reached.

C. Estimation of *p-values* with Monte Carlo cross validation method. We propose to model the *p-value* empirically as a data's size dependent function ($p(n)$) by Monte Carlo cross validation (MCCV) with replacement (3). This way, the effect of the sample bias on the *p-value* can be ignored. The procedure followed for the estimation of the *p-values* is illustrated in Figure S1 and the corresponding pseudocode in Algorithm 1. Notice that in Algorithm 1 we estimate $p(n)$ in two different ways, using

Algorithm 1 *p-value* estimation

```
N_max ← min{|S_A|, |(S_B|}

𝒩 ← exp(grid[log(n_0), log(n_∞), gridsize])
𝒩 ← int(𝒩)

a ← N_max/k_1 n_0
b ← k_2 N_max/n_∞
ℱ ← exp(grid[log(a), log(b), gridsize])
ℱ ← int(ℱ)

for i in 0 : length(𝒩) do
    n_i ← 𝒩[i]
    # Start Monte Carlo cross validation:
    for f = 0 : ℱ[i] do
        s_A ← sample(S_A, n_i)
        s_B ← sample(S_B, n_i)
        𝒫_i ← save the p-value of test(s_A, s_B)
    𝒫 ← save mean(𝒫_i)
    𝒫̄ ← save mean(𝒫_i)
p_L ← LOWESS(𝒫̄)
p_e ← exponential.fit(𝒫)
```

either a locally weighted scatter plot smoothing (LOWESS) approximation (4) (p_L) or and exponential fit (p_e). The main reason to do this is that we use a standard curve fitting (LOWESS) to show that $p(n)$ is exponential.

Aiming to compare two sets of values, S_A and S_B, and to determine if there exists statistically significant differences between them, the estimation of the *p-values* is done in pairs (i.e., two sub-samples are compared each time). A range of values needs to be defined for both the sample size and the number of folds in MCCV. These are given by the grids $\mathcal{N}$ and $\mathcal{F}$, respectively.

The range for all possible sub-sample sizes (n) goes from 2 (n_0) to the smallest size between samples S_A and S_B (N_{max}). A grid covering all these values for large N_{max}, is computationally expensive and redundant. As the *p-value* tends to zero when $n \to \infty$, the most important information is condensed in the smallest values of n. So the grid $\mathcal{N}$ follows an exponential distribution from $n_0 = 2$. Similarly, a large enough upper-limit (n_∞) is chosen such that it ensures a fast computation ($n_\infty \ll N_{max}$) and the convergence to zero of $p(n)$. Hence, $\mathcal{N}$ is determined as

$$\mathcal{N} = \{n_i : n_i \in \exp\left(\mathcal{U}\left(\log(n_0), \log(n_\infty)\right)\right)\}, \qquad [8]$$

where $\mathcal{U}$ is the uniform distribution that goes from $\log(n_0)$ to $log(n_\infty)$. In MCCV, the number of folds can be extremely large when working with large datasets and a small partition. On the contrary, for a large partition size, the number of folds might decrease dramatically. To compensate for both situations, $\mathcal{F}$ is defined as given below

$$\mathcal{F} = \{f_i : f_i \in \exp\left(\mathcal{U}\left(a, b\right)\right)\}, \text{ where } a = \log\left(\frac{N_{max}}{k_1 n_0}\right), b = \log\left(\frac{k_2 N_{max}}{n_\infty}\right), \qquad [9]$$

$\mathcal{U}$ is the uniform distribution, k_1 controls the upper-limit on the number of folds for small sub-sample sizes, and k_2 controls the lower-limit for large sub-sample sizes. Note that the number of elements in $\mathcal{N}$ and $\mathcal{F}$ are the same.

Finally, for each n_i in $\mathcal{N}$, MCCV is applied to obtain the set of *p-values* defined as

$$\mathcal{P}_i = \left\{p_i^j, j \in [1, ..., f_i]\right\}, \qquad [10]$$

being f_i the number of folds in $\mathcal{F}$ that corresponds with the sub-sample size n_i in $\mathcal{N}$.

D. Assessment of minimum data size needed for statistical significance (n_α). The estimation of the *p-value* function $p(n)$ supports the computation of the minimum data size needed to obtain statistically significant differences (n_α). This value is the solution to the equation

$$\alpha = ae^{-cn_\alpha}. \qquad [11]$$

As explained in the online methods, the parameters a and c are the result of fitting an exponential function to the empirical values $\mathcal{P}_i$ in Equation 10. Therefore, there exists an intrinsic bias in the estimated values a and c. Additionally, the estimation of $p(n)$, and specially, its decay (parameter c) can be less precise when the data size is small. (See Figure 2*d* in the main manuscript), so n_α in Equation 11 can be biased.

For this reason, the calculation of a more conservative estimator, $\hat{n}_\alpha$, is strongly recommended

$$\hat{n}_\alpha = arg \min_{n_i} \left\{\left(\overline{p_i} + \sigma_{\overline{p_i}}\right) < \alpha\right\}, \qquad [12]$$

where $\overline{p_i}$ represents the mean of $\mathcal{P}_i$ and $\sigma_{\overline{p_i}}$, the mean standard error (SEM), which is included to correct for the variability of the estimated *p-values*. Hence, the estimator of the theoretical value will always be slightly larger

$$\hat{n}_\alpha \geq n_\alpha.$$

However, $\hat{n}_\alpha$ can only be provided when the sample is large enough to cover those n values smaller or equal to $\hat{n}_\alpha$. For this reason, whenever the data is not large enough, the theoretical value n_α in Equation 11 is also given. In these cases, even if n_α and $\Theta_{\alpha,\gamma}$ might be slightly deviated, they still serve as an indicator of the existence of statistical significance.

2. Experimental data

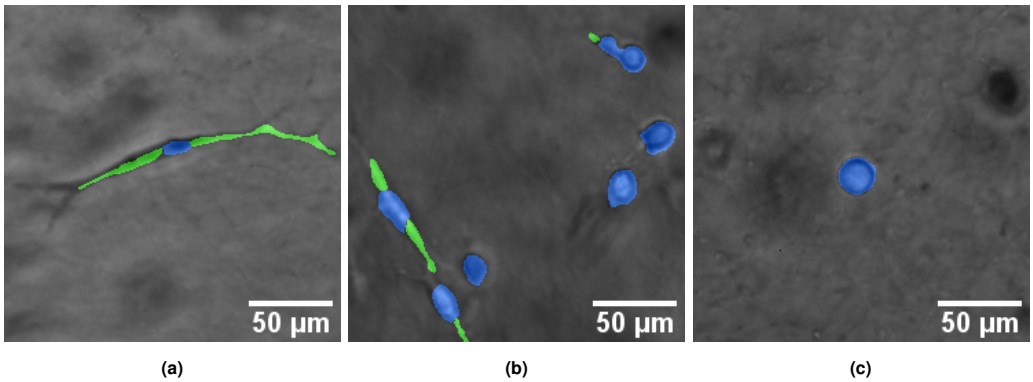

Fig. S2. Segmentation of phase contrast microscopy images of cancer cells (MDA-MB-231) embedded in a 3D collagen Type I matrix. Cell bodies are labeled in blue and cell protrusions in green. Images of (a) control cells and cells treated at (b) 1 nM and (c) 50 nM Taxol were acquired with a 10 x magnification objective.

A. Experiment 1: Drug analysis on phase contrast microscopy data. Image processing analysis provided the necessary information to distinguish the cellular body and protrusions of each cell in the videos (Figure S2). Hence, we got eight different measurements: cell body size (C_S), cell body perimeter (C_P), cell body roundness (C_R), cell with at least one protrusion (P_b), protrusion size (P_S), protrusion perimeter (P_P), protrusion length (P_L) and protrusion diameter (P_D). Each of the morphological measurements is given in microns. Table S1 contains the complete list of variables.

Table S1. List of computed variables. C: continuous variable. B: binary variable.

Cell body			Cell protrusions		
Feature	**Name**	**Type**	**Feature**	**Name**	**Type**
Area (μm^2)	C_S	Categorical	Area (μm^2)	P_S	Categorical
Perimeter (μm)	C_P	Categorical	Perimeter (μm)	P_P	Categorical
Roundness	C_R	Categorical	Length (μm)	P_L	Categorical
Protrusions	P_b	Binary	Diameter (μm)	P_D	Categorical

In Figure S3, the distribution of the the variables used in the analysis of cellular shape is shown. The cellular body changes with the amount of Taxol used to treat cells. When they are treated at 50 nM Taxol, the cellular body is biger and more rounded (Figure S3a). Besides, this same treatment prevents cells from producing long and thick protrusions (Figure S3b).

None of the continuous variables presented in Table S1 follows a normal distribution, so the comparison was carried out by the Mann-Whitney U-test (1). P_b was a binary variable (Figure S3c) to distinguish protruding cells (value one). Therefore, it was analyzed by means of Pearson χ^2-test for categorical data (5).

As per the number of observations reported in Table S2 and following methodology guidelines, we set $\mathcal{P}_i$ with $n_0 = 2$ and $n_\infty = 2500$. $\mathcal{N}$ and $\mathcal{F}$ were set to have 190 points. The number of folds $\mathcal{F}$ described in the Supplementary Material, was computed using $k_1 = 1$, $k_2 = 20$ and $N_{max} = 11037$. These values were chosen to have a reasonable number of permutations for both small and large sample sizes (6.000 permutations when $n_0 = 2$, and 90 when $n_{190} = 2500$, respectively). Table S3 contains the estimated coefficients a and c of the exponential curve (ae^{-cn}) for each of the variables we analyzed and each pair of comparisons (Control - 1 nM Taxol, Control - 50 nM Taxol, and 1 nM - 50 nM Taxol).

Figures S4, S5 and S6 show the shape of each of the exponential curves that result with the coefficients in Table S3. To determine whether Taxol has a significant effect in cell's morphology, $\Theta_{\alpha,\gamma}$ was chosen such that $\alpha = 0.05$ (95% of statistical significance) and $\gamma = 5 \cdot 10^{-6}$.

When comparing the control group and 1 nM Taxol, there are not statistically significant differences in cell body morphology: the curve $p(n)$ of any cell body feature decreases slowly, i.e. $\hat{n}_\alpha$ and n_γ are large and $\Theta_{\alpha,\gamma} = 0$ (Table S3 and Figure S4a). On the other hand, cells at 50 nM Taxol have a significantly higher roundness index and bigger cellular body: when comparing

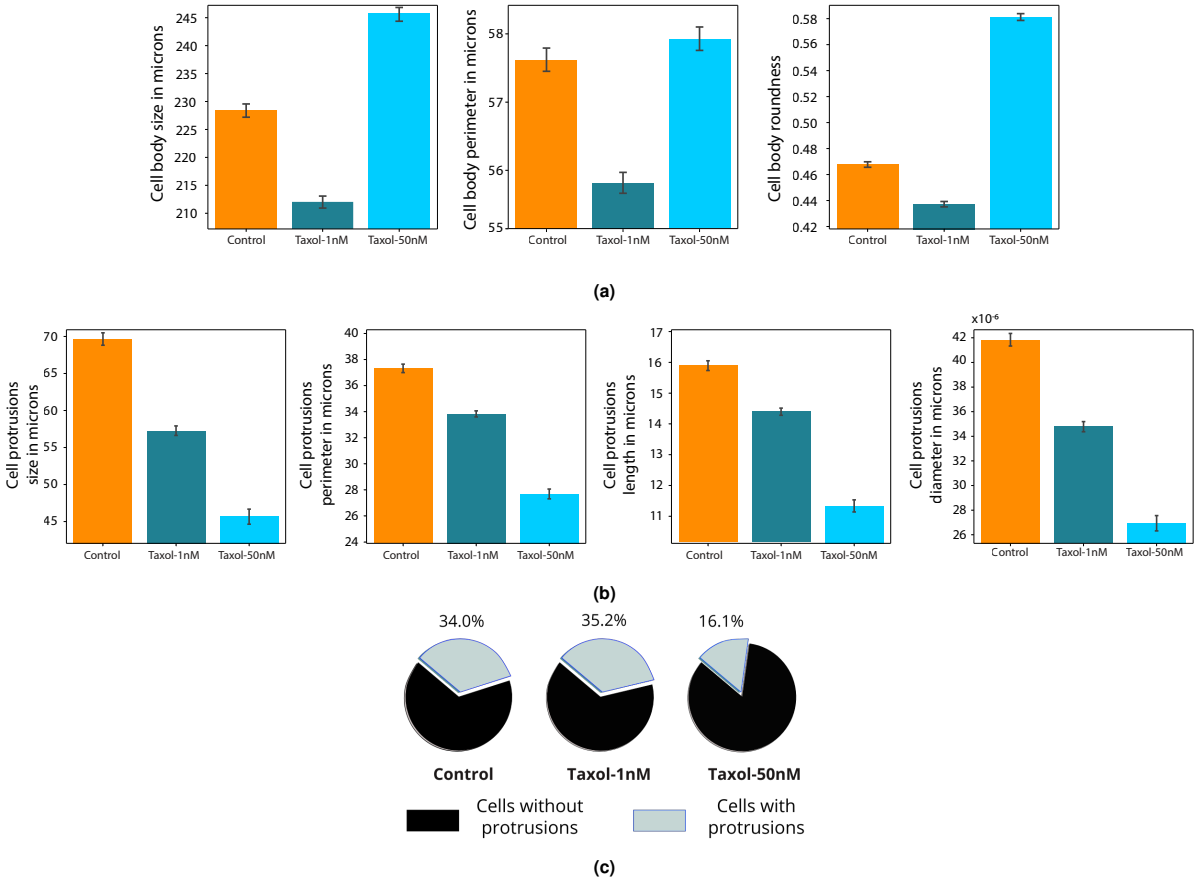

Fig. S3. Quantitative variables used to measure (a) cell bodies and (b) cellular protrusions morphology, and (c) the ratio of cells with and without protrusions for the three different treatment groups (control, 1 nM Taxol, 50 nM Taxol). Error bars in (a) and (b) correspond to the confidence interval at 99 %

Treatment group	Cell body	Cell protrusions
Control	77,700	45,871
1 nM Taxol	74,713	42,798
50 nM Taxol	46,162	11,037

Table S2. Number of observations (cell body and their protrusions -if present-) per treatment group.

control vs. 50 nM Taxol or 1 nM vs. 50 nM Taxol the curves corresponding to C_R and C_S decrease rapidly, i.e. $\hat{n}_\alpha$ and n_γ are small, and $\Theta_{\alpha,\gamma} = 1$ (Table S3 and Figures S4b and S4c, respectively). For C_P, it is also possible to appreciate some differences when comparing 1 nM with 50 nM Taxol group, i.e. $\Theta_{\alpha,\gamma} = 1$ (Table S3). Namely, the blue curve shown in Figure S4c decreases faster than those in Figures S4a and S4b.

Similar results are obtained when the morphology of cellular protrusions is evaluated (Table S3 and Figure S5). While Taxol at 1 nM does not change their morphology ($\Theta_{\alpha,\gamma} = 0$ in Table S3 and Figure S5a), the effect of Taxol at 50 nM is much larger ($\Theta_{\alpha,\gamma} = 1$ in Table S3, Figures S4b and S4c) .

Usually, when a categorical variable such as P_b is analyzed, the input of a statistical test is a percentage rather than the raw data. Hence, when there is no statistical significance, the $p(n)$ function shoots up, as for instance in Figure **??**. However, when there exist statistical differences, $p(n)$ decreases and it is possible to analyze its decay, as in Figures **??** and **??**. With all, we can say that the formation of protrusions is inhibited when 50 nM Taxol are administered: there is a significant reduction in the number of cells that form protrusions and their protrusions are smaller (shorter and thinner) (S3, Figures S3b and S3c).

B. Experiment 2: Drug analysis on flow cytometry data. Flow cytometry is a technique that generates a large amount of data for each experiment. Consequently, any statistical test for groups comparison results in a vanishing *p-value*. To avoid that situation, practitioners tend to reduce the data to a single, representative measure for subject. For instance, Khoury *et al.* (6) acquired fluorescence intensity data from 6 different subjects and compute the median fluorescence intensity (MFI) for each of them. So, the statistical test is just applied on the 6 MFI values. However, our proposal of estimating the *p-value* as a function of the sample size enables to incorporate in the test the information given by the whole dataset and take into consideration the deviation and bias present in the data.

Variables	Cell body size (C_S)				Cell body perimeter (C_P)				Cell body roundness (C_R)				Cell with protrusions (P_b)			
Comparison	a	c	$\hat{n}_\alpha$	$\Theta_{\alpha,\gamma}$	a	c	$\hat{n}_\alpha$	$\Theta_{\alpha,\gamma}$	a	c	$\hat{n}_\alpha$	$\Theta_{\alpha,\gamma}$	a	c	$\hat{n}_\alpha$	$\Theta_{\alpha,\gamma}$
C - 1T	0.258	0.0026	670	**0**	0.258	0.0017	1160	**0**	0.259	0.0029	617	**0**	0.435	-0.0005	∞	**0**
C - 50T	0.263	0.0075	250	1	0.256	0.0014	1331	**0**	0.282	0.0400	47	1	0.198	0.0345	42	1
1T - 50T	0.272	0.0216	83	1	0.264	0.0072	257	1	0.292	0.0648	29	1	0.195	0.0351	41	1
Variables	Protrusions size (P_S)				Protrusions perimeter (P_P)				Protrusions length (P_L)				Protrusions diameter (P_D)			
Comparison	a	c	$\hat{n}_\alpha$	$\Theta_{\alpha,\gamma}$	a	c	$\hat{n}_\alpha$	$\Theta_{\alpha,\gamma}$	a	c	$\hat{n}_\alpha$	$\Theta_{\alpha,\gamma}$	a	c	$\hat{n}_\alpha$	$\Theta_{\alpha,\gamma}$
C - 1T	0.250	0.0031	563	1	0.248	0.0019	754	**0**	0.251	0.0011	1695	**0**	0.251	0.0023	707	**0**
C - 50T	0.246	0.0221	75	1	0.241	0.0276	58	1	0.250	0.0289	58	1	0.250	0.0248	68	1
1T - 50T	0.250	0.0100	170	1	0.256	0.0175	98	1	0.255	0.0211	80	1	0.247	0.0134	127	1

Table S3. Parameters of the exponential function ae^{-cn} and estimated minimum size $\hat{n}_\alpha$ for each of the analyzed variables. C: control, 1T: 1 nM Taxol and 50T: 50 nM Taxol. $a \in [0.195, 0.435]$, $c \in \left[-5 \cdot 10^{-4}, 0.0648\right]$, $\alpha = 0.05$ **and** $\gamma = 5 \cdot 10^{-6}$.

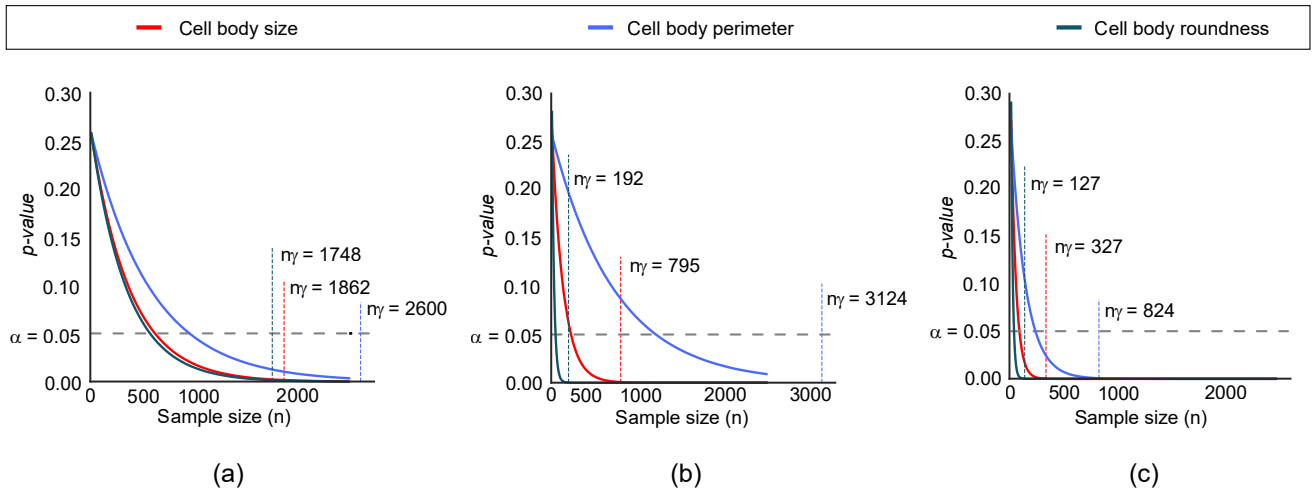

Fig. S4. Results obtained for cell body morphology when control, 1 nM Taxol and 50 nM Taxol groups are compared as (a) control vs. 1 nM Taxol, (b) control vs. 50 nM Taxol and (c) 1 nM vs. 50 nM Taxol. Vertical lines correspond to the convergence point n_γ with $\gamma = 5 \cdot 10^{-6}$.

To illustrate the proposed procedure, we analyzed the flow cytometry data provided in (7) to determine the transcriptional changes induced by the *in vivo* exposure of human eosinophils to glucocorticoids. Khoury *et al.* (6) studied eosinophil surface proteins after being exposed to glucocorticoids and demonstrated that this exposure causes the apoptosis of human eosinophils (eosinopenia) once they migrate out of the blood circulation.

While they performed an extensive analysis, we have focused our study on the data related to the chemokine receptor gene CXCR4. In particular, the expression of CXCR4 on the surface of human eosinophils after being exposed for 2 hours to vehicle, 20 mcg/dL and 200 mcg/dL of Methylprednisolone (MP). After filtering the raw data to discard noise and debris, we got clean distributions to analyze (Figure 2b, right). The *p-value* curves computed for pair group comparisons were the result of applying Mann-Whitney U statistical tests following the proposed procedure. Then, the exponential curves (ae^{-cn}) were fitted (Figure 2b, left). The parameters configuration was $n_0 = 2$, $\mathcal{N}$ and $\mathcal{F}$ had 200 points, $k_1 = 1$, $k_2 = 20$, $\alpha = 0.05$, and $\gamma = 5 \cdot 10^{-6}$. As the number of data points was lower than $1,000$, n_∞ and N_{max} where chosen to be the minimum number of points for each group pair being compared. The results are summarized in Table S4. Our results are similar to those in (6), in the sense that we also find a differential expression of CXCR4 when eosinophils are exposed to glucocorticoids.

C. Experiment 3: Cellular age characterization by means of biomolecular and biophysical properties. The data extracted from Phillip *et al.* in (8) was analysed with the following parameter configuration: for each of the stated variables, the data belonging to the group of 2 years-old was compared with the data from $\{3, 9, 16, 29, 35, 55, 65, 85, 92\}$ and $\{3, 9, 16, 29, 35, 55, 65, 85, 96\}$ years-old human donors to test cell motility and morphology, respectively. For each pair of groups, the distribution of the n-dependent *p-values* was obtained using the Mann-Whitney U statistical test. Then, the parameters of the exponential function (ae^{-cn}) were fitted. The parameter configuration was $n_0 = 2$, $\mathcal{N}$ and $\mathcal{F}$ had 200 points, $k_1 = 1$, $k_2 = 20$, $\alpha = 0.05$, and $\gamma = 5 \cdot 10^{-6}$. As the number of data points was lower than 1000, n_∞ and N_{max} where chosen to be the minimum number of points of each pair of groups being compared. The results for cell motility and cell morphology are summarized in Tables S5 and S6, respectively.

Measures	CXCR4 surface expression				
Comparison	a	c	$\hat{n}_\alpha$	n_α	$\Theta_{\alpha,\gamma}$
Vehicle - MP 20 mcg/dL	0.286	0.048	40	36	1
Vehicle - MP 200 mcg/dL	0.286	0.049	37	35	1
MP 20 mcg/dL - MP 200 mcg/dL	0.256	$1.69 \cdot 10^{-4}$	-	9680	0

Table S4. Parameters of the exponential function ae^{-cn} for the differential expression of CXCR4, theoretical minimum size (n_α) and its estimator ($\hat{n}_\alpha$) for a 95% ($\alpha = 0.05$) of statistical significance, and decision index $\Theta_{\alpha,\gamma}$, for $\gamma = 5 \cdot 10^{-6}$. MP: Methylprednisolone

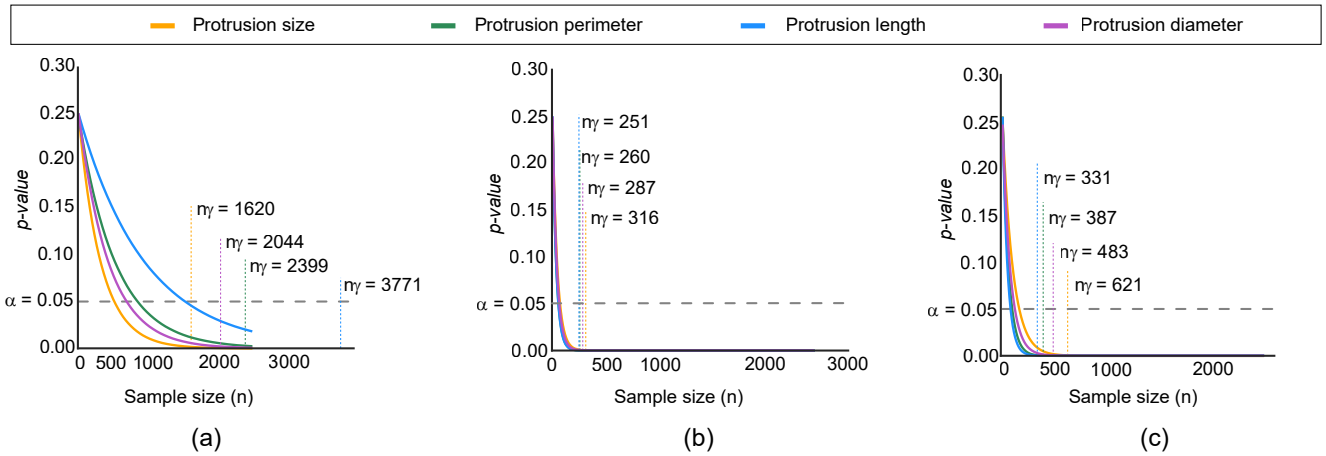

Fig. S5. Graphical illustration of the results obtained for cell protrusions morphology when (a) control and 1 nM Taxol, (b) control and 50 nM Taxol and (c) 1 nM Taxol and 50 nM Taxol groups are compared. Vertical lines correspond to the convergence point n_γ with $\gamma = 5 \cdot 10^{-6}$.

3. Test of robustness

The variability in the statistical significance of the results caused by the selection of the parameter γ and the grid sizes $\mathcal{N}$ and $\mathcal{F}$ are characterized in this section. The method is first tested using theoretical distributions and then, using the real data from Experiment 1.

A. Test of robustness on theoretical data. We simulated normal distributions to test the method in a theoretical scenario: $\mathcal{N}(0,1)$ was compared with $\mathcal{N}(0.01,1)$, $\mathcal{N}(0.1,1)$, $\mathcal{N}(0.25,1)$, $\mathcal{N}(0.5,1)$, $\mathcal{N}(0.75,1)$, $\mathcal{N}(1,1)$, $\mathcal{N}(2,1)$ and $\mathcal{N}(3,1)$. For the most similar cases such as $\mathcal{N}(0,1)$ vs. $\mathcal{N}(0.01,1)$, or $\mathcal{N}(0,1)$ vs. $\mathcal{N}(0.1,1)$, it is expected to obtain $\Theta_{\alpha,\gamma} = 0$. While for the most different distributions such as $\mathcal{N}(0,1)$ vs. $\mathcal{N}(2,1)$, or $\mathcal{N}(0,1)$ vs. $\mathcal{N}(3,1)$, $\Theta_{\alpha,\gamma} = 1$.

Theoretically, an optimal grid $\mathcal{N}$ would be the one that covers the values from $n_0 = 2$ to $n_\infty = N_{max}$. This set up entails an extremely large amount of computations, while it suffices a value $n_\infty \approx 1000$ to understand what is the tendency of the data. If $p(n)$ converges to zero when $n > 1,000$, then it can be assumed that $p(n)$ does not represent a statistical significant case. Hence, $n_\infty = 2,500$ is large enough for the implementation of the method. As the *p-values* for very small samples are especially unstable and small samples are not representative of any real scenario, the minimum value n_0 can be increased. The number of permutations for each n_i can be decreased as well: while the amount of data to analyze may be infinite, it is enough to study a certain limited number of different data subsamples to approach a realistic scenario. Hence, to test the robustness of the proposed method, we set grids $\mathcal{N}$ and $\mathcal{F}$ using $n_0 = 20$, $n_\infty = 2,500$, $N_{max} = 10,000$, $k_1 = 1$ and $k_2 = 20$ in Equations 8 and 9, respectively. Both $\mathcal{N}$ and $\mathcal{F}$ were configured to have a size of 200 points. Thus, MCCV is repeated 200 times. See Figure S1 for the workflow.

With this grid parameters, for γ in the set $\{2.5 \cdot 10^{-6}, 5 \cdot 10^{-6}, 5 \cdot 10^{-5}, 5 \cdot 10^{-4}\}$, we run the pipeline to evaluate the effect of γ value on the rejection of the null hypothesis of the Mann-Whitney U statistical test (Table S7). The results obtained help to assess the most suitable γ value. Specifically, the decision about γ relies on the result obtained for the comparison between $\mathcal{N}(0,1)$ and $\mathcal{N}(0.25,1)$: while the distance $\delta_{\alpha,\gamma}$ for $\gamma = 5 \cdot 10^{-6}$ and $\gamma = 5 \cdot 10^{-5}$ expresses the same ($\delta_{\alpha,\gamma} = \pm 5.84$), the minimum data size needed to observe statistically significance differences is low enough as to reject the null hypothesis, i.e. $\hat{n}_\alpha = 186$ and $\Theta_{\alpha,\gamma} = 1$. Hence, the value chosen for the following simulations and for the real data is $\gamma = 5 \cdot 10^{-6}$. (Table S7).

To test the computational limitations of the method, we evaluated the value $\Theta_{0.05,5\cdot 10^{-6}}$ reducing $\mathcal{N}$ and $\mathcal{F}$: $\mathcal{N}$ was chosen to be a grid of size 10, 20, 50, 100, 150 or 200 points and the values in $\mathcal{F}$ were reduced by a factor of 1/2, 1/3, 1/5 and 1/10 (i.e., each of the values in the original $\mathcal{F}$ was multiplied by this fraction). The experiment was repeated 100 times on each of the setups, so the probability of obtaining exactly the same $\Theta_{0.05,5\cdot 10^{-6}}$ (Table S7) and the stability of the method could be evaluated. The information given in Table S8 lets the assessment of (1) the size of $\mathcal{N}$ and (2) the number of folds in $\mathcal{F}$. In most cases, the probability obtained was 100%, which shows that the final results are very stable. When $N(0,1)$ and $N(0.5,1)$ were compared with small grid parameters, this probability decreased slightly to $89 - 96\%$ (Table S8). In conclusion, the number of computations could be considerably reduced, for the example, to $\mathcal{N} = 50$ and $\mathcal{F} = 0.2\mathcal{F}$).

B. Test of robustness on real data.

B.1. The* p-value *can be estimated by an exponential function.. Repeating the procedure followed with simulation of normal distributions, we verify that the condition for $p(n)$ being exponential is satisfied again: in Figures S7a-c all LOWESS fittings have exponential shapes, and in Figures S7d-f the quotient $p'(n)/p(n)$ of LOWESS fits are constant.

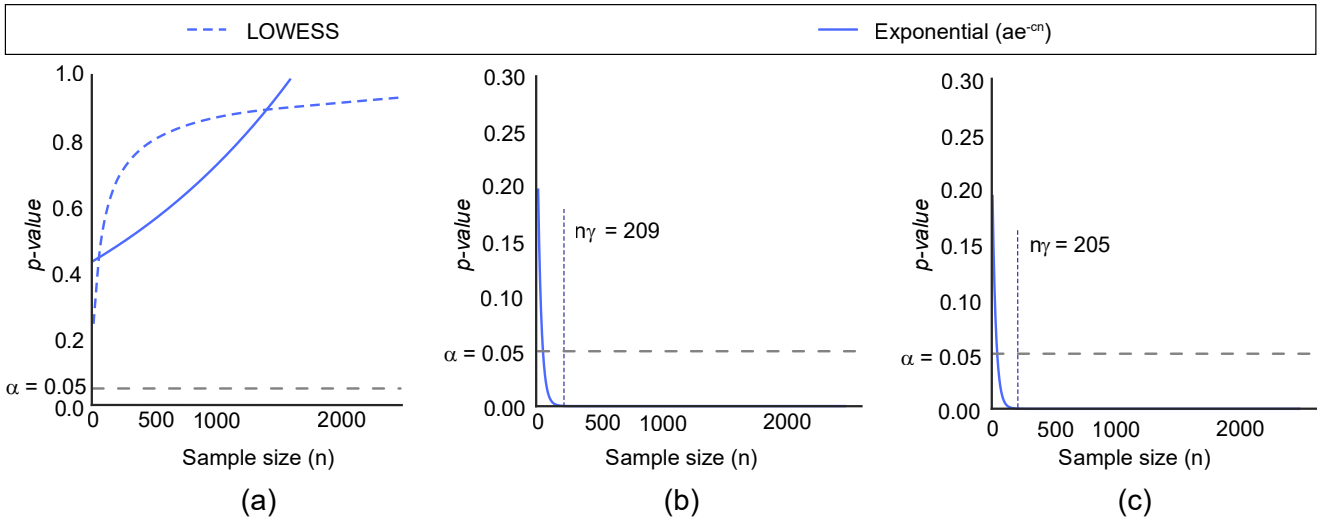

Fig. S6. Estimation of $p(n)$ with χ^2 test for contingency tables when comparing cells (not) having at least one protrusion (binary variable cell with protrusions: P_b) for (a) control and 1 nM Taxol groups, (b) control and 50 nM Taxol groups, and (c) 1 nM and 50 nM Taxol groups. In the leftmost plot, both locally weighted scatter plot smoothing (LOWESS) and exponential fit of estimated *p-values* are shown. Vertical lines correspond to the convergence point n_γ with $\gamma = 5 \cdot 10^{-6}$.

B.2. Robustness of the convergence threshold and required computational load. The distribution of real data is more complex than the typical Gaussian distributions due to the presence of noise, large deviation of the data or leverage points. Following the same procedure as in Section A, we tested the reliability of the proposed method using the data we extracted from the microscopy images (Experiment 1, main manuscript). We evaluated both, the effect of varying the convergence threshold γ and the required computational load. Looking at Tables S3 and S9, it can be appreciated once again that $\gamma = 5 \cdot 10^{-6}$ is a good value for the convergence threshold. Smaller values of γ result in the rejection of the null hypothesis for cases in which $\hat{n}_\alpha > 1000$ as cellular protrusions length. Similarly, when $\gamma = 5 \cdot 10^{-5}$, there are cases as cell body roundness for which $\hat{n}_\alpha < 100$ and $\Theta_{\alpha,\gamma} = 0$. Therefore, once again, $\gamma = 5 \cdot 10^{-6}$ seems to be an appropriate value to measure statistical significance at $\alpha = 0.05$ significance level (Table S9).

When the grid $\mathcal{N}$ is large enough and $\mathcal{F}$ has large numbers, the result is completely stable (Table S10). However, when these values are dramatically reduced (for example, $\mathcal{N} = 10$, $\mathcal{F} = 0.02\mathcal{F}_0$, $\mathcal{F} = 0.01\mathcal{F}_0$), the reproducibility of the results may degrade. This fact is specially evident when the variables are noisy as in the case of those that measure cellular protrusions morphology (P_S, P_P, P_L and P_D), being the noisier the protrusions size and perimeter. In summary, while large grid parameters ensure stable results, with the information given in Tables S8 and S10, seems reasonable to reduce the number of computations to $\mathcal{N} \geq 50$ and $\mathcal{F} \geq 0.2\mathcal{F}$.

Measures	MSD-6min					MSD-60 min					Persistence primary axis				
Comparison	a	c	$\hat{n}_\alpha$	n_α	$\Theta_{\alpha,\gamma}$	a	c	$\hat{n}_\alpha$		$\Theta_{\alpha,\gamma}$	a	c	$\hat{n}_\alpha$	n_α	$\Theta_{\alpha,\gamma}$
A02 - A03	0.296	0.038	46	46	1	0.276	0.009	-	187	1	0.287	0.012	-	145	1
A02 - A09	0.270	0.001	-	1333	0	0.295	0.000	-	$4.72 \cdot 10^{17}$	0	0.276	0.002	-	896	0
A02 - A16	0.289	0.014	-	125	1	0.262	0.001	-	1955	0	0.310	0.084	23	21	1
A02 - A29	0.394	0.217	12	9	1	0.357	0.161	13	12	1	0.287	0.013	-	137	1
A02 - A35	0.315	0.155	12	11	1	0.335	0.123	18	15	1	0.274	0.002	-	752	0
A02 - A55	0.286	0.018	-	99	1	0.283	0.032	52	54	1	0.286	0.000	-	$4.99 \cdot 10^{15}$	0
A02 - A65	0.306	0.032	48	56	1	0.345	0.113	15	17	1	0.330	0.097	19	19	1
A02 - A85	0.282	0.008	-	229	1	0.303	0.040	46	45	1	0.315	0.050	34	36	1
A02 - A92	0.345	0.189	12	10	1	0.386	0.228	10	8	1	0.331	0.079	27	24	1
Measures	Persistence secondary axis					Diffusivity primary axis					Diffusivity secondary axis				
Comparison	a	c	$\hat{n}_\alpha$	n_α	$\Theta_{\alpha,\gamma}$	a	c	$\hat{n}_\alpha$	n_α	$\Theta_{\alpha,\gamma}$	a	c	$\hat{n}_\alpha$	n_α	$\Theta_{\alpha,\gamma}$
A02 - A03	0.276	0.001	-	1399	0	0.301	0.000	-	1.47E+17	0	0.290	0.034	52	51	1
A02 - A09	0.266	0.001	-	3023	0	0.275	0.006	-	289	1	0.279	0.051	32	33	1
A02 - A16	0.295	0.069	27	25	1	0.293	0.030	56	58	1	0.268	0.004	-	443	0
A02 - A29	0.301	0.019	-	94	1	0.335	0.078	27	24	1	0.354	0.160	15	12	1
A02 - A35	0.281	0.014	-	126	1	0.312	0.061	30	29	1	0.312	0.117	19	15	1
A02 - A55	0.284	0.015	-	116	1	0.285	0.016	-	109	1	0.284	0.008	-	213	1
A02 - A65	0.294	0.030	-	58	1	0.327	0.086	24	21	1	0.312	0.100	15	18	1
A02 - A85	0.287	0.027	54	65	1	0.304	0.051	33	35	1	0.269	0.004	-	397	1
A02 - A92	0.293	0.016	-	112	1	0.380	0.217	11	9	1	0.295	0.025	-	71	1
Measures	Total diffusivity					Anisotropy									
Comparison	a	c	$\hat{n}_\alpha$	n_α	$\Theta_{\alpha,\gamma}$	a	c	$\hat{n}_\alpha$	n_α	$\Theta_{\alpha,\gamma}$					
A02 - A03	0.284	0.000	-	$8.70 \cdot 10^{18}$	0	0.291	0.024	69	72	1					
A02 - A09	0.280	0.000	-	$1.68 \cdot 10^{19}$	0	0.307	0.075	26	24	1					
A02 - A16	0.297	0.036	40	48	1	0.293	0.000	-	$1.16 \cdot 10^{15}$	0					
A02 - A29	0.360	0.122	17	16	1	0.290	0.016	-	106	1					
A02 - A35	0.330	0.093	22	20	1	0.281	0.014	-	120	1					
A02 - A55	0.276	0.006	-	269	1	0.309	0.043	40	42	1					
A02 - A65	0.330	0.105	19	18	1	0.266	0.000	-	35475	0					
A02 - A85	0.306	0.052	33	34	1	0.293	0.015	-	120	1					
A02 - A92	0.379	0.201	11	10	1	0.359	0.104	20	19	1					

Table S5. Parameters of the exponential function ae^{-cn} for cell motility, theoretical minimum size (n_α) and estimated one ($\hat{n}_\alpha$) for a 95% ($\alpha = 0.05$) of statistical significance, and decision index $\Theta_{\alpha,\gamma}$, for $\gamma = 5 \cdot 10^{-6}$.

References

1. Mann HB, Whitney DR (1947) On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann. Math. Stat.* 18(1):50–60.
2. Student (1908) The Probable Error of a Mean. *Biometrika* 6(1):1.
3. Xu QS, Liang YZ (2001) Monte Carlo cross validation. *Chemom. Intell. Lab. Syst.* 56(1):1–11.
4. Cleveland WS (1979) Robust Locally Weighted Regression and Smoothing Scatterplots. *J. Am. Stat. Assoc.* 74(368):829–836.
5. Pearson K (1900) X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *London, Edinburgh, Dublin Philos. Mag. J. Sci.* 50(302):157–175.
6. Khoury P, et al. (2018) Glucocorticoid-induced eosinopenia in humans can be linked to early transcriptional events. *Allergy* 73(10):2076–2079.
7. Gadkari M, et al. (2018) Transcript- and protein-level analyses of the response of human eosinophils to glucocorticoids. *Sci. Data* 5:180275.
8. Phillip JM, et al. (2017) Biophysical and biomolecular determination of cellular age in humans. *Nat. Biomed. Eng.* 1(7):0093.

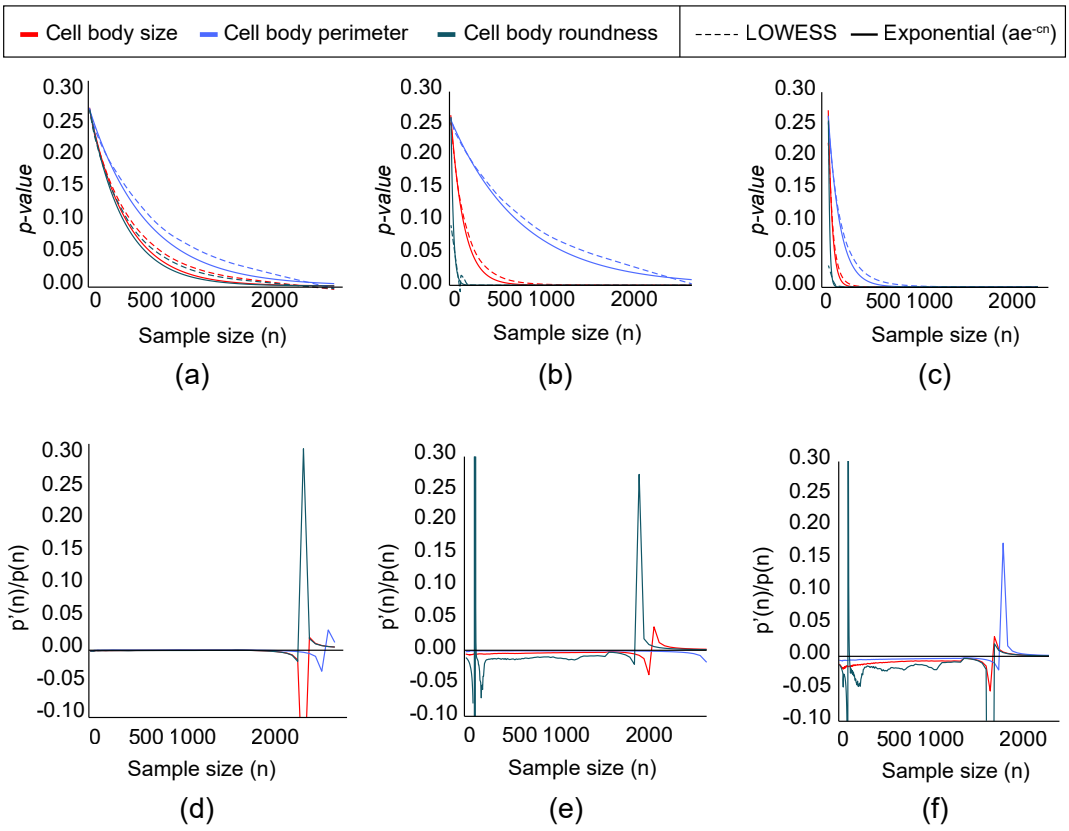

Fig. S7. Curve fitting. Three different simulations of $p(n)$ are computed with the data obtained from the experimental test. Cell body size, perimeter and roundness of three different cell groups where compared by means of the Mann-Whitney U statistical test and Monte Carlo cross validation: (a) and (d) control cells versus cells treated with Taxol at 1 nM; (b) and (e) control cells versus cells treated with Taxol at 50 nM and (c) and (f) cells treated with Taxol at 1 nM and 50 nM. To each of the mean *p-value* sets (i.e., the output of Monte Carlo cross validation), a locally weighted scatter plot smoothing (LOWESS) (4) curve was fit to get the initial shape of $p(n)$. Likewise, an exponential function was fit to all the *p-values* obtained in each fold of the Monte Carlo cross validation (before averaging). Both the LOWESS and exponential curves are shown in (a), (b) and (c). The quotient between each LOWESS curve and its differential is shown in (d), (e) and (f). The constant quotients and the accurate exponential fits show empirically that the $p(n)$ functions have an exponential nature.

Measures	Size (pixels)					Perimeter (pixels)					Long axis length (pixels)				
Comparison	a	c	$\hat{n}_\alpha$	n_α	$\Theta_{\alpha,\gamma}$	a	c	$\hat{n}_\alpha$	n_α	$\Theta_{\alpha,\gamma}$	a	c	$\hat{n}_\alpha$	n_α	$\Theta_{\alpha,\gamma}$
A02 - A03	0.267	0.002	509	688	0	0.271	0.009	203	183	1	0.276	0.022	76	76	1
A02 - A09	0.264	0.002	-	923	0	0.260	0.000	-	22578	0	0.274	0.008	-	201	1
A02 - A16	0.320	0.149	13	12	1	0.340	0.210	10	9	1	0.477	0.459	6	4	1
A02 - A29	0.312	0.129	16	14	1	0.313	0.121	16	15	1	0.325	0.144	14	12	1
A02 - A35	0.311	0.106	19	17	1	0.295	0.084	21	21	1	0.316	0.133	16	13	1
A02 - A55	0.313	0.112	18	16	1	0.323	0.136	13	13	1	0.364	0.243	10	8	1
A02 - A65	0.326	0.137	16	13	1	0.373	0.270	8	7	1	0.401	0.335	7	6	1
A02 - A85	0.330	0.167	13	11	1	0.431	0.401	6	5	1	0.442	0.403	6	5	1
A02 - A96	0.351	0.252	9	7	1	0.370	0.320	7	6	1	0.422	0.421	5	5	1
Measures	Short axis length (pixels)					Orientation					Solidity				
Comparison	a	c	$\hat{n}_\alpha$	n_α	$\Theta_{\alpha,\gamma}$	a	c	$\hat{n}_\alpha$	n_α	$\Theta_{\alpha,\gamma}$	a	c	$\hat{n}_\alpha$	n_α	$\Theta_{\alpha,\gamma}$
A02 - A03	0.267	0.005	284	330	1	0.265	0.000	-	$2.38 \cdot 10^{18}$	0	0.273	0.013	124	126	1
A02 - A09	0.263	0.000	-	3801	0	0.267	0.000	-	$1.09 \cdot 10^{16}$	0	0.280	0.026	69	65	1
A02 - A16	0.264	0.005	321	345	1	0.262	0.001	-	$1.41 \cdot 10^{03}$	0	0.266	0.003	386	485	0
A02 - A29	0.284	0.040	48	43	1	0.262	0.000	-	$1.53 \cdot 10^{14}$	0	0.276	0.025	74	68	1
A02 - A35	0.272	0.011	159	160	1	0.265	0.000	-	$4.60 \cdot 10^{17}$	0	0.318	0.125	16	14	1
A02 - A55	0.260	0.000	-	6616	0	0.265	0.000	-	$1.65 \cdot 10^{20}$	0	0.267	0.005	267	319	1
A02 - A65	0.291	0.057	34	30	1	0.262	0.001	-	$3.20 \cdot 10^{03}$	0	0.275	0.015	117	112	1
A02 - A85	0.311	0.103	20	17	1	0.265	0.000	-	$8.27 \cdot 10^{16}$	0	0.321	0.177	11	10	1
A02 - A96	0.295	0.069	27	25	1	0.261	0.000	-	$1.13 \cdot 10^{16}$	0	0.298	0.077	26	23	1
Measures	Equivalent diameter (pixels)					Aspect ratio					Circularity				
Comparison	a	c	$\hat{n}_\alpha$	n_α	$\Theta_{\alpha,\gamma}$	a	c	$\hat{n}_\alpha$	n_α	$\Theta_{\alpha,\gamma}$	a	c	$\hat{n}_\alpha$	n_α	$\Theta_{\alpha,\gamma}$
A02 - A03	0.267	0.002	509	688	0	0.282	0.027	74	63	1	0.277	0.022	78	77	1
A02 - A09	0.264	0.002	-	923	0	0.277	0.016	103	104	1	0.261	0.000	-	$1.33 \cdot 10^{13}$	0
A02 - A16	0.320	0.149	13	12	1	0.384	0.327	7	6	1	0.319	0.162	12	11	1
A02 - A29	0.312	0.129	16	14	1	0.275	0.019	86	90	1	0.260	0.001	-	2313	0
A02 - A35	0.311	0.106	19	17	1	0.281	0.042	44	40	1	0.259	0.000	-	$1.22 \cdot 10^{15}$	0
A02 - A55	0.313	0.112	18	16	1	0.318	0.154	13	12	1	0.289	0.056	36	31	1
A02 - A65	0.326	0.137	16	13	1	0.306	0.094	20	19	1	0.313	0.113	18	16	1
A02 - A85	0.330	0.167	13	11	1	0.290	0.068	27	25	1	0.338	0.238	9	8	1
A02 - A96	0.351	0.252	9	7	1	0.326	0.149	13	12	1	0.337	0.187	11	10	1
Measures	Roundness														
Comparison	a	c	$\hat{n}_\alpha$	n_α	$\Theta_{\alpha,\gamma}$										
A02 - A03	0.277	0.022	78	77	1										
A02 - A09	0.261	0.000	-	$1.33 \cdot 10^{13}$	0										
A02 - A16	0.319	0.162	12	11	1										
A02 - A29	0.260	0.001	-	2313	0										
A02 - A35	0.259	0.000	-	$1.2 \cdot 10^{15}$	0										
A02 - A55	0.289	0.056	36	31	1										
A02 - A65	0.313	0.113	18	16	1										
A02 - A85	0.338	0.238	9	8	1										
A02 - A96	0.337	0.187	11	10	1										

Table S6. Parameters of the exponential function ae^{-cn} for cell nuclei morphology measures, theoretical minimum size (n_α) and estimated one ($\hat{n}_\alpha$) for a 95% ($\alpha = 0.05$) of statistical significance, and decision index $\Theta_{\alpha,\gamma}$, for $\gamma = 5 \cdot 10^{-6}$.

γ	$2.5 \cdot 10^{-6}$		$5 \cdot 10^{-6}$		$5 \cdot 10^{-5}$		$5 \cdot 10^{-4}$		
Comparison	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\hat{n}_\alpha$
$\mathcal{N}(0,1)$ - $\mathcal{N}(0,1)$	0	-490.25	0	-490.25	0	-490.25	0	-490.25	0
$\mathcal{N}(0,1)$ - $\mathcal{N}(0.01,1)$	0	-3576	0	-488.25	0	-488.25	0	-488.25	0
$\mathcal{N}(0,1)$ - $\mathcal{N}(0.1,1)$	1	0.782	0	-21.922	0	-79.938	0	-42.375	1237
$\mathcal{N}(0,1)$ - $\mathcal{N}(0.25,1)$	1	9.523	1	5.848	0	-5.840	0	-12.945	186
$\mathcal{N}(0,1)$ - $\mathcal{N}(0.5,1)$	1	0	1	0	1	0	1	0	45
$\mathcal{N}(0,1)$ - $\mathcal{N}(0.75,1)$	1	0	1	0	1	0	1	0	22
$\mathcal{N}(0,1)$ - $\mathcal{N}(1,1)$	1	0	1	0	1	0	1	0	0
$\mathcal{N}(0,1)$ - $\mathcal{N}(1.5,1)$	1	0	1	0	1	0	1	0	0
$\mathcal{N}(0,1)$ - $\mathcal{N}(2,1)$	1	0	1	0	1	0	1	0	0
$\mathcal{N}(0,1)$ - $\mathcal{N}(2.5,1)$	1	0	1	0	1	0	1	0	0
$\mathcal{N}(0,1)$ - $\mathcal{N}(3,1)$	1	0	1	0	1	0	1	0	0

Table S7. Table of decision index $\Theta_{\alpha,\gamma}$, difference $\delta_{\alpha,\gamma} = A_{\alpha_\gamma} - A_{p(n_\gamma)}$ and estimated minimum size ($\hat{n}_\alpha$) for $\alpha = 0.05$ and $\gamma = 2.5 \cdot 10^{-6}, 5 \cdot 10^{-6}, 5 \cdot 10^{-5}, 5 \cdot 10^{-4}$.

		$\mathcal{N}$ grid size								$\mathcal{N}$ grid size					
Comparison	$\mathcal{F}$ reduction	10	20	50	100	150	200	Comparison	$\mathcal{F}$ reduction	10	20	50	100	150	200
$\mathcal{N}(0,1)$ - $\mathcal{N}(0,1)$	0.1	100	100	100	100	100	100	$\mathcal{N}(0,1)$ - $\mathcal{N}(0.75,1)$	0.1	100	100	100	100	100	100
	0.2	100	100	100	100	100	100		0.2	100	100	100	100	100	100
	1/3	100	100	100	100	100	100		1/3	100	100	100	100	100	100
	0.5	100	100	100	100	100	100		0.5	100	100	100	100	100	100
	1	100	100	100	100	100	100		1	100	100	100	100	100	100
$\mathcal{N}(0,1)$ - $\mathcal{N}(0.01,1)$	0.1	100	100	100	100	100	100	$\mathcal{N}(0,1)$ - $\mathcal{N}(1,1)$	0.1	100	100	100	100	100	100
	0.2	100	100	100	100	100	100		0.2	100	100	100	100	100	100
	1/3	100	100	100	100	100	100		1/3	100	100	100	100	100	100
	0.5	100	100	100	100	100	100		0.5	100	100	100	100	100	100
	1	100	100	100	100	100	100		1	100	100	100	100	100	100
$\mathcal{N}(0,1)$ - $\mathcal{N}(0.1,1)$	0.1	99	100	100	100	100	100	$\mathcal{N}(0,1)$ - $\mathcal{N}(1.5,1)$	0.1	100	100	100	100	100	100
	0.2	100	100	100	100	100	100		0.2	100	100	100	100	100	100
	1/3	100	100	100	100	100	100		1/3	100	100	100	100	100	100
	0.5	100	100	100	100	100	100		0.5	100	100	100	100	100	100
	1	100	100	100	100	100	100		1	100	100	100	100	100	100
$\mathcal{N}(0,1)$ - $\mathcal{N}(0.25,1)$	0.1	100	100	100	100	100	100	$\mathcal{N}(0,1)$ - $\mathcal{N}(2,1)$	0.1	100	100	100	100	100	100
	0.2	100	100	100	100	100	100		0.2	100	100	100	100	100	100
	1/3	100	100	100	100	100	100		1/3	100	100	100	100	100	100
	0.5	100	100	100	100	100	100		0.5	100	100	100	100	100	100
	1	100	100	100	100	100	100		1	100	100	100	100	100	100
$\mathcal{N}(0,1)$ - $\mathcal{N}(0.5,1)$	0.1	89	95	99	99	100	100	$\mathcal{N}(0,1)$ - $\mathcal{N}(3,1)$	0.1	100	100	100	100	100	100
	0.2	96	99	100	100	100	100		0.2	100	100	100	100	100	100
	1/3	100	99	100	100	100	100		1/3	100	100	100	100	100	100
	0.5	99	100	100	100	100	100		0.5	100	100	100	100	100	100
	1	100	100	100	100	100	100		1	100	100	100	100	100	100

Table S8. Table of results for different sizes of $\mathcal{N}$ and $\mathcal{F}$ grids. Each value represents the probability (%) of obtaining the same decision index $\Theta_{0.05,5}$ as the one shown in Table S9.

Variable	Cell body size								
γ	$2.5 \cdot 10^{-6}$		$5 \cdot 10^{-6}$		$5 \cdot 10^{-5}$		$5 \cdot 10^{-4}$		
Comparison	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\hat{n}_\alpha$
Control - 1 nM Taxol	1	8.67	0	-4.12	0	-41.26	-	-	670
Control - 50 nM Taxol	1	9.37	1	4.82	0	-9.69	-	-	250
1 nM - 50 nM Taxol	1	5.38	1	3.74	0	-1.47	0	5.84	83
Variable	Cell body perimeter								
γ	$2.5 \cdot 10^{-6}$		$5 \cdot 10^{-6}$		$5 \cdot 10^{-5}$		$5 \cdot 10^{-4}$		
Comparison	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\hat{n}_\alpha$
Control - 1 nM Taxol	1	1.37	0	-17.84	0	-69.41	-	-	1160
Control - 50 nM Taxol	0	-5.75	0	-29.94	0	-90.28	-	-	1331
1 nM - 50 nM Taxol	1	9.44	1	4.69	0	-10.43	-	-	257
Variable	Cell body roundness								
γ	$2.5 \cdot 10^{-6}$		$5 \cdot 10^{-6}$		$5 \cdot 10^{-5}$		$5 \cdot 10^{-4}$		
Comparison	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\hat{n}_\alpha$
Control - 1 nM Taxol	1	9.17	0	-2.67	0	-37.41	-	-	617
Control - 50 nM Taxol	1	3.42	1	2.57	0	-0.25	0	-2.87	47
1 nM - 50 nM Taxol	1	2.39	1	1.84	1	0.05	0	-1.59	29
Variable	Protrusions binary								
γ	$2.5 \cdot 10^{-6}$		$5 \cdot 10^{-6}$		$5 \cdot 10^{-5}$		$5 \cdot 10^{-4}$		
Comparison	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\hat{n}_\alpha$
Control - 1 nM Taxol	0	0.00	0	0.00	0	0.00	-	-	-
Control - 50 nM Taxol	1	5.72	1	4.72	1	1.41	0	-1.55	42
1 nM - 50 nM Taxol	1	5.72	1	4.72	1	1.46	0	-1.42	41
Variable	Protrusions size								
γ	$2.5 \cdot 10^{-6}$		$5 \cdot 10^{-6}$		$5 \cdot 10^{-5}$		$5 \cdot 10^{-4}$		
Comparison	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\hat{n}_\alpha$
Control - 1 nM Taxol	1	11.99	1	1.15	0	-31.16	-	-	563
Control - 50 nM Taxol	1	6.23	1	4.68	0	-0.42	0	-4.73	75
1 nM - 50 nM Taxol	1	9.50	1	6.07	0	-4.98	0	-11.98	170
Variable	Protrusions perimeter								
γ	$2.5 \cdot 10^{-6}$		$5 \cdot 10^{-6}$		$5 \cdot 10^{-5}$		$5 \cdot 10^{-4}$		
Comparison	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\hat{n}_\alpha$
Control - 1 nM Taxol	1	7.75	0	-9.85	0	-58.09	-	-	754
Control - 50 nM Taxol	1	5.54	1	4.29	1	0.15	0	-3.40	58
1 nM - 50 nM Taxol	1	6.77	1	4.77	0	-1.63	0	-6.71	98
Variable	Protrusions length								
γ	$2.5 \cdot 10^{-6}$		$5 \cdot 10^{-6}$		$5 \cdot 10^{-5}$		$5 \cdot 10^{-4}$		
Comparison	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\hat{n}_\alpha$
Control - 1 nM Taxol	0	-14.70	0	-45.38	0	-114.17	-	-	1695
Control - 50 nM Taxol	1	5.10	1	3.90	0	-0.04	0	-3.45	58
1 nM - 50 nM Taxol	1	6.09	1	4.45	0	-0.90	0	-5.37	80
Variable	Protrusions diameter								
γ	$2.5 \cdot 10^{-6}$		$5 \cdot 10^{-6}$		$5 \cdot 10^{-5}$		$5 \cdot 10^{-4}$		
Comparison	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\Theta_{\alpha,\gamma}$	$\delta_{\alpha,\gamma}$	$\hat{n}_\alpha$
Control - 1 nM Taxol	1	9.64	0	-4.80	0	-45.90	0	-12.24	707
Control - 50 nM Taxol	1	5.63	1	4.29	0	-0.29	0	-4.20	68
1 nM - 50 nM Taxol	1	8.34	1	5.81	0	-2.49	0	-8.57	127

Table S9. Table of decision index $\Theta_{\alpha,\gamma}$, difference $\delta_{\alpha,\gamma} = A_{\alpha_\gamma} - A_{p(n_\gamma)}$ and estimated minimum size $\hat{n}_\alpha$ for $\alpha = 0.05$ and $\gamma = 2.5 \cdot 10^{-6}, 5 \cdot 10^{-6}, 5 \cdot 10^{-5}, 5 \cdot 10^{-4}$.

Variables		Cell body size				Cell body perimeter				Cell body roundness				Protrusions binary			
		$\mathcal{N}$ grid size				$\mathcal{N}$ grid size				$\mathcal{N}$ grid size				$\mathcal{N}$ grid size			
Comparison	reduction	10	20	50	100	10	20	50	100	10	20	50	100	10	20	50	100
Control - 1 nM Taxol	0.01	80	98	82	94	95	100	99	100	71	92	71	79	100	100	100	100
	0.02	86	99	79	99	99	100	100	100	71	98	84	95	100	100	100	100
	0.1	100	100	99	100	100	100	100	100	93	100	98	100	100	100	100	100
	0.2	100	100	100	100	100	100	100	100	96	100	100	100	100	100	100	100
	1/3	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	0.5	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Control - 50 nM Taxol	0.01	93	100	99	100	100	100	100	100	89	100	93	94	92	99	88	87
	0.02	100	100	100	100	100	100	100	100	100	100	96	100	94	99	87	98
	0.1	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	0.2	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	1/3	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	0.5	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
1 nM - 50 nM Taxol	0.01	95	99	98	98	98	100	100	100	80	94	81	89	79	98	78	91
	0.02	98	100	100	100	99	100	100	100	89	99	93	96	90	99	90	99
	0.1	100	100	100	100	100	100	100	100	100	100	100	100	100	100	98	100
	0.2	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	1/3	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	0.5	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Variables		Protrusions size				Protrusions perimeter				Protrusions length				Protrusions diameter			
		n grid size				n grid size				n grid size				n grid size			
Comparison	reduction	10	20	50	100	10	20	50	100	10	20	50		10	20	50	100
Control - 1 nM Taxol	0.01	45	28	46	51	76	79	80	82	88	84	97	98	76	84	76	77
	0.02	36	35	50	60	74	65	79	90	86	89	99	100	78	79	85	90
	0.1	51	36	61	71	78	58	98	100	98	98	100	100	86	82	97	99
	0.2	64	24	59	83	89	57	100	100	99	100	100	100	91	95	100	100
	1/3	68	22	75	97	97	68	100	100	100	100	100	100	98	98	100	100
	0.5	66	17	82	99	98	75	100	100	100	100	100	100	100	100	100	100
Control - 50 nM Taxol	0.01	76	79	91	95	82	73	84	99	74	69	88	93	74	67	90	97
	0.02	91	98	99	99	85	85	95	98	75	89	93	98	83	86	94	98
	0.1	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	0.2	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	1/3	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	0.5	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
1 nM - 50 nM Taxol	0.01	71	67	96	99	73	67	92	100	72	75	90	95	75	72	94	100
	0.02	92	80	100	100	87	84	99	100	94	90	98	100	86	87	99	100
	0.1	100	100	100	100	98	100	100	100	99	100	100	100	100	100	100	100
	0.2	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	1/3	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	0.5	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

Table S10. Table of results for different sizes of $\mathcal{N}$ and $\mathcal{F}$ grids. Each value represents the probability (%) of obtaining the same decision index $\Theta_{0.05,5e-06}$ as the one shown in Table S9.