WORKING PAPER SERIES

P-hacking in Clinical Trials: A Meta-Analytical Approach

Norman Belas/Paul Bengart/Bodo Vogt

Working Paper No. 19/2017

OTTO VON GUERICKE
UNIVERSITÄT
MAGDEBURG

FACULTY OF ECONOMICS
AND MANAGEMENT

Impressum (§ 5 TMG)

Herausgeber:
Otto-von-Guericke-Universität Magdeburg
Fakultät für Wirtschaftswissenschaft
Der Dekan

Verantwortlich für diese Ausgabe:
Norman Belas, Paul Bengart and Bodo Vogt
Otto-von-Guericke-Universität Magdeburg
Fakultät für Wirtschaftswissenschaft
Postfach 4120
39016 Magdeburg
Germany

http://www.fww.ovgu.de/femm

Bezug über den Herausgeber
ISSN 1615-4274

P-hacking in Clinical Trials: A Meta-Analytical Approach

Norman Belas* Paul Bengart† Bodo Vogt‡

December 7, 2017

Abstract

Clinical trials play a decisive role in the drug approval processes. By completing a p-curve analysis of a newly compiled data set that consists of thousands of clinical trials, we substantiate that the occurrence of p-hacking in clinical trials is not merely hypothetical. Medical and pharmaceutical research consists of both primary and secondary study endpoints. The primary finding covers the main effect, which directly influences the approval process, while the secondary outcome delivers further additional information. For primary p-curves, we observed an abnormal increase in the p-value frequency at common significance thresholds, while the secondary p-curves exhibited no such anomaly.

Keywords: P-hacking, Publication Bias, Reporting Bias, Clinical Trials

JEL Codes: C18, H51

*Otto-von-Guericke-University Magdeburg, Faculty of Economics and Management, Chair in Empirical Economics & Medical Faculty, Institute of Social Medicine and Health Economics, Postfach 4120, 39016 Magdeburg, Germany

†Otto-von-Guericke-University Magdeburg, Faculty of Economics and Management, Chair in Empirical Economics

‡Otto-von-Guericke-University Magdeburg, Faculty of Economics and Management, Chair in Empirical Economics & Medical Faculty, Institute of Social Medicine and Health Economics

1 Introduction

The reproducibility problem is increasingly attracting attention within the area of academic research. Over the last decade, a debate has evolved that spans various scientific disciplines and concerns the misuse of statistical inference that originates from a misunderstanding of fundamental statistical concepts or conscious deception (Ioannidis, 2005; Simmons et al., 2011; Wasserstein & Lazar, 2016; Farland et al., 2016). This aberration leads to the erroneous processing of data and misinterpretation of empirical findings. The overall lack of corroboration derails scientific progress because it engenders a situation in which researchers linger on in the twilight of selective reporting. In 2016, the American Statistical Association released a statement on p-value interpretation and application in the best sense of the wisdom: cum hoc ergo propter hoc. The statement emphasizes how satisfying significance thresholds does not imply a higher probability of true hypotheses; vice versa, the p-value only indicates how incompatible the data are with the hypothesis (Wasserstein & Lazar, 2016; Krzywinski & Altman, 2017). The lack of good statistical practice results in the need for researchers to enhance the reproducibility of scientific findings in a variety of academic fields, including clinical research (Ioannidis & Trikalinos, 2007; Ioannidis et al., 2009; Prinz et al., 2011; Collins et al., 2014). In this paper, we present a meta-analytical approach that proves that the misuse of statistical methods does occur in clinical trials.

The main objectives of the regulatory authorities that oversee the pharmaceutical industry are to protect and support public health by monitoring the development and distribution of prescription and nonprescription medication. Clinical trials are an essential component of the process by which drugs are approved for use in humans (Seife, 2015). Statistical validity is the baseline for ethical clinical research (Emanuel et al. 2000). Scientifically unsound research not only violates ethical standards (CIOMS, 2002), it can also have negative consequences for human health since an insufficient efficacy of drugs might lead to slower or less significant recovery. Furthermore, any unanticipated side effects of the drug could negatively impact the

quality of life of the patient. However, at present, supervisions focus is mostly on technical and medical procedures rather than on statistical approaches (Bradshaw, 2009). Even if it is assumed that data dredging in clinical trials occurs very rarely (Al-Marzouki et al., 2005; Buyse, M. et al., 1999), statistical analyses incorporates various parameters that allow for the manipulation, deception, and modification of the underlying data.

To investigate the gap between expectations and reality in terms of the application of good clinical practice, we engage the meta-analytical p-curve approach (Simonsohn et al., 2014), which is prominent in psychological science, to evaluate a specially compiled data set that consists of thousands of clinical trials conducted in the United States over the last 15 years. Main question is: What can the p-value distribution in the body of clinical studies tell us about whether there was data-dredging?

The p-curve provides an opportunity to distinguish between selective reporting or specification search on the one hand, and the truth on the other hand (Simonsohn et al., 2014). It is an observation about the frequency distribution of p-values. In the current study, this distribution was shaped by the results of clinical trials. The rationale that underpins the p-curve is simple: If there is no effect, the p-curve has a uniform distribution. If effects in the studies occur, then there is a right-skewed distribution of the p-curve. The more statistical power the steeper the slope of the p-curve becomes. In other words: A right-skewed p-curve, which encompasses a set of independent findings with continuously decreasing p-values from low to high, is an indicator of evidential value. When p-curves differ from that ideal-typical shape, we assume a partial lack of evidential value in the set of findings. The intensity of data dredging is determined by an abnormality in the shape of the p-curves, respectively the deviation from the ideal-typical graph. Power law frequency examples, like Zipfs law or the Pareto distribution, span many scientific fields and also appear in empirical data (Newman, 2005). The p-curve itself is a power law probability distribution that is applicable to meta-analyses of any empirical scientific discipline. Strategies by which data dredging can be identified via scrutinizing probability values have evolved over recent years within several

research fields (Masicampo & Lalande, 2012; Jager & Leek, 2014; Head et al., 2015). While this approach is by no means new, its use as a means of detecting p-hacking in clinical trials is.

Drug approval processes can span over a decade and the expenditure pharmaceutical companies invest in the process of conducting clinical trials and gaining approval can escalate into billions of U.S. dollars (Scannell et al., 2012). However, drugs can generate average annual peak sales of around one billion U.S. dollars (Mullard, 2014). The higher industry expenditures translate into higher expectations for future returns. A remarkable possibility of an unsuccessful approval process remains (DiMasi et al., 2003) such that investments can easily become sunk costs. Naturally, the main players in the pharmaceutical industry attempt to increase the probability of a new drug gaining approval and take proactive action to contemporaneously improve the prospect of its success on the drugs market. Contract research organizations and research institutes conduct time-consuming clinical trials on behalf of pharmaceutical companies. In addition to fostering a positive reputation, researchers are also heavily focused on having their findings published in professional journals (Weir & Murray, 2011). In light of the congruent incentive structures that motivate stakeholders, industry and research organizations, and given the fact that human health is the pivot of all efforts, an empirical meta-analysis of clinical trials appears to be long overdue.

2 Data & Method

The data set that is assessed in this study was compiled on the basis of clinical trials that were conducted to gain Food and Drug Administration (FDA) approval for new drugs. The results of the clinical trials were provided by the U.S. National Institutes of Health (NIH). The FDA approval process constitutes a regulatory market entry barrier for pharmaceuticals. When the drug has passed preclinical trials in laboratory animal testing, and the FDA accepts sponsor-companies Investigational New Drug Application (IND), tests on humans

can begin. Our analysis takes into account all clinical trials that were conducted before market maturity between 2002 and December 2016. Only studies that completed phase three were considered since this phase is the last before beginning the new drug application process (NDA). The acquired data is from 6,081 studies with results, of which 2,841 provided at least one p-value. We only employ exact p-values, e.g., those without relational operators, to preclude the reason for the occurrence of any peak at the five percent level being due to inaccurately reported p-values. The final data set consisted of 1,177 completed trials, 20 percent of which the 19,584 p-values were related to primary outcomes and 80 percent to the secondary. Primary study endpoints tackle the major effect of the drug, while secondary study endpoints capture additional information for explanatory purposes (D'Agostino, 2000; Meinert, 2012). Single trials include several measurements related to the same medication but differ in terms of dose or duration of treatment. To achieve the statistical independence of p-values and trials, we drew a random sample without replacement, capturing a single p-value for each trial.

Despite ex-ante deception capabilities during data generating processes, p-hacking focuses on ex-post researcher degrees of freedom (Simmons et al., 2011). It appears widely in various scientific disciplines including medical and health research (Head et al., 2015). P-hacking comprises ex-post determination of sample size, excluding outliers, or a specification search concerning various numbers of variables or alternating covariates (Simmons et al., 2011; Simonsohn et al., 2014). Making use of researchers degrees of freedom might be perceived to represent a trivial offense; however, there is a thin line between obvious scientific deception and fraud; although the appearance of fraud is assumed to be less frequent (Fanelli, 2009).

The FDAs strict guidelines for reporting and the persistent monitoring process has led to the disclosure of many cases in which clinical research standards have been violated (Seife, 2015; Buyse, 1999; George & Buyse, 2015). The existence of additional undiscovered violation of good clinical practice is very likely even if the amount of further irregularities is difficult to quantify (Seife, 2015; George & Buyse, 2015). The distinct verification of p-hacking

in single studies is impossible without replication. However, reproducibility comes to the fore in various disciplines; e.g., in the replicability projects that are common in the field of psychology (Open Science Collaboration, 2015) or the sound replication initiatives in experimental economics (Camerer et al., 2016). Literature that supports the imperative character of reproducibility in medical science also exists. Earlier studies addressed the statistical validation of fraud (Ranstam et al., 2000) or evaluated data on clinical trials by industry researchers, with one paper indicating that literature data on potential drug targets should be viewed with caution (Prinz et al., 2011). Additionally, previous research has presented evidence that indicates misconduct occurs more frequently in medical and pharmacological research than in other disciplines (Fanelli, 2009). The publication bias, which is commonly known as the file-drawer effect, describes how the residual scientific work above the five percent level ends up in the file drawer. This is also a serious problem because it entails there is less tolerance of null results and a greater willingness to publish only significant results (Rosenthal, 1979).

First, and contrary to replication of single studies, the meta-analytical approach allows for an investigation that evaluates the overall statistical validity of clinical results. Second, the data could help to avoid the file-drawer problem since FDA rules specify that all non-significant results should be made publically available (NIH, 2016). However, the problem remains that many industry and research organizations often fail to publish the results of clinical trials. This represents a violation of ethical standards, can have potential ramifications for patient well-being, and prevents quality improvements by auditioning (Saito & Gill, 2014; Miller et al., 2015; Powell-Smith & Goldacre, 2016). To avoid this, the U.S. Department of Health and Human Services (HHS), as the mother agency of the NIH and FDA, has implemented rules that specify the amount of information about clinical trials that should be provided to the public (NIH, 2016).

The figures in this paper cover expected and observed p-curves. An expected p-curve depicts an ideal-typical form of probability intervals of observing certain p-values. Its calculation

depends on the parameters of sample size, effect size, and non-centrality. Hence, its shape bases on the level of statistical power following the function: The higher the power, the more pronounced the p-curves right-skew is (Simonsohn et al., 2014). The second step involves the depiction of observed p-curves regarding primary and secondary endpoints. We subdivide the continuous frequency distribution into clusters to enable the statistical analysis of all p-curves. We then examine the p-value intervals at the ten percent significance level. The resulting p-curves allow us to observe the frequency distribution of the p-values. We employ two different types of proper and bounded intervals. The first interval is closed, $[.0, .01] = \{p_i \mid .0 \leq p_i \leq .01\}$ denoted after the right interval boundary: .01 interval. Subsequent intervals to the ten percent level are all left-open and right-closed, $(p_l,\ p_r] = \{p_i \mid p_l \leq p_i \leq p_r\}$ also denoted after the right interval boundary: p_r-interval.

Our first hypothesis states that the primary and secondary outcomes would be inhomogeneous interval frequencies. We assume that significant inhomogeneous interval frequencies, respectively p-curve shapes, could be attributed to differences in the relevance for approval and publication chance. We test the stochastic independence of primary and secondary p-curves using a chi-squared test on the frequency of p-value intervals. The next hypothesis analyzes anomalies in the p-curve. According to literature, we assume that a significant increase in the frequency of p-values slightly below the five percent level could be assessed as evidence of p-hacking (Bradshaw 2009; Simonsohn et al., 2014). The graph of a p-curve reporting evidential values asymptotically approaches zero. The higher the p-value, the lower is the respective interval frequency. To ensure drug approval and enhance publication chance, many industry and research organizations deliver results that are below the five percent significance level at a minimum. Our second hypothesis states that, if p-values are more likely to occur at the five percent level than in the preceding interval, this represents evidence of p-hacking in clinical trials. We provide two proper and bounded intervals to consider that various boundaries, indeed, affect the statistical results of binomial tests differently. The lower small interval is closed and has following properties:

$[.04, .045] = \{p_i \mid .04 \leq p_i \leq .045\}$. The upper small interval is left-open and right-closed with the following properties: $(.045, .05] = \{p_i \mid .045 \leq p_i \leq .05\}$. The secondly tested intervals are larger. The lower large interval is closed and has the following properties: $[.03, .04] = \{p_i \mid .03 \leq p_i \leq .04\}$. The upper large interval is left-open and right-closed with the following properties: $(.04, .05] = \{p_i \mid .04 \leq p_i \leq .05\}$. To test our second hypothesis, we perform binomial tests on the small and large interval pairs including the five percent threshold and the preceding interval, taking into consideration both study endpoints.

3 Results

The discrete representation of the observed p-values reveals major differences between the primary and secondary distribution, as exhibited in Figure 1. As this figure displays, the secondary p-value frequency corresponds to the shape of an expected p-curve. The primary p-values indicates strong anomalies in the distribution, which exhibits three solitaire peaks, above the first one near the null, rising at three significance thresholds. The most prominent peak rises at the most prominent significance level of five percent, while the two remaining p-value peaks rise at the .01 and .03 thresholds. The .01 threshold is popular in most scientific disciplines. The secondary p-values in Figure 1 do not exhibit such abnormalities.

Although its discrete representation has a fortuitous nature, the secondary p-values roughly correspond to the hypothetical expectations, which are confirmed by visual examination of Figure 1. A prominent peak in the shape of the primary p-curve can be observed in Figure 2 at the .05 interval. The frequency of p-values occurring at the five percent level is higher than the relative frequency of the .02 , .03 and .04 intervals. Interestingly, a harsh decline immediately after the .05 interval becomes evident since both the .06 and .07 intervals have a lower relative frequency than the .08 and .09 intervals. This anomaly disappears not before the ten percent significance level. The inconsistency becomes even more obvious

if we compare the shape of the primary and secondary p-curves in figure 2 on page 9. The secondary p-curve exhibits a monotonously decreasing function over almost all p-value intervals. As the secondary p-curve in figure 2 corresponds to the ideal-typical shape of the expected p-curve, the primary p-curve contradicts it.

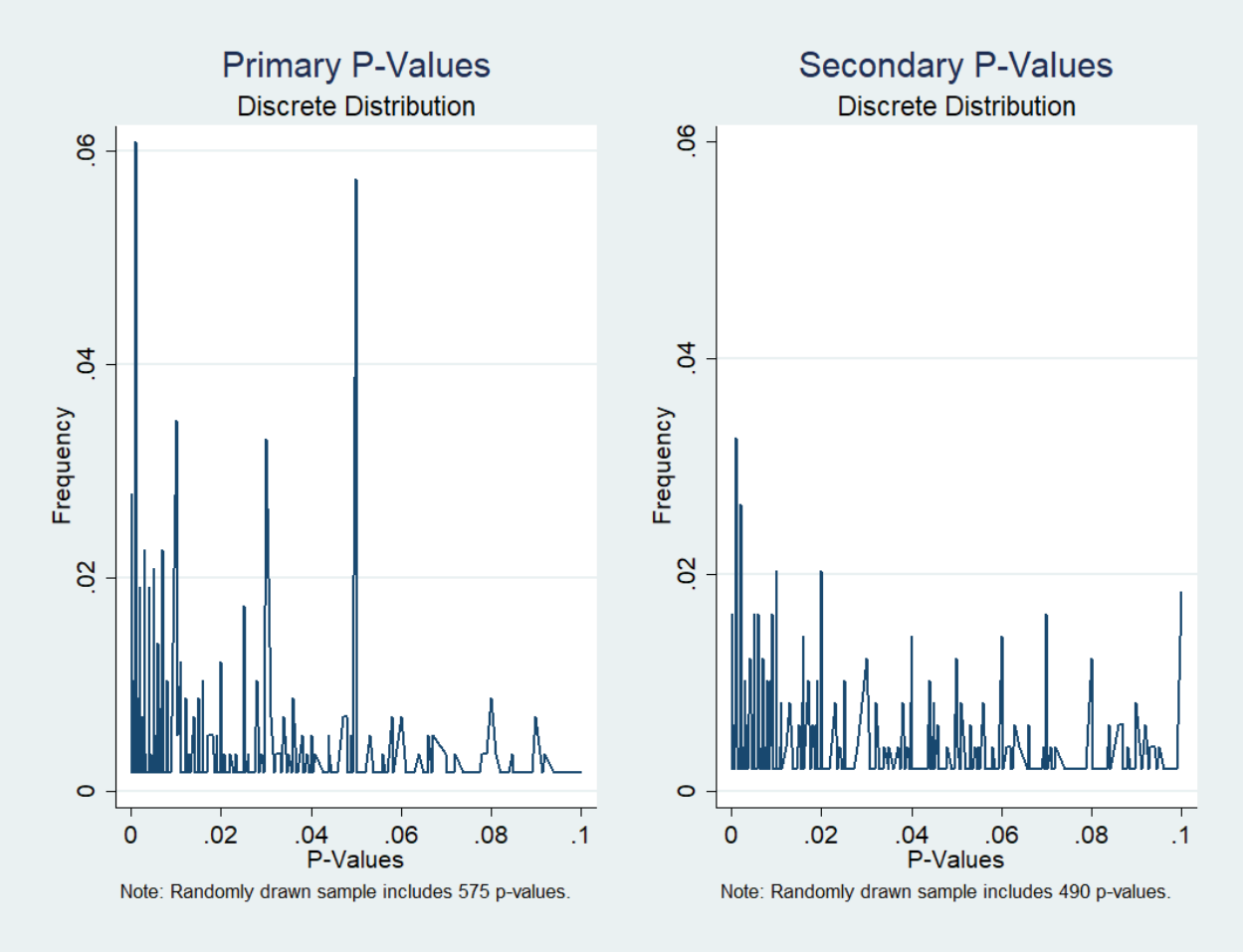

Figure 1: Discrete representation of observed p-values from primary study endpoints to the ten percent significance level. The abscissa displays the p-values while the ordinate depicts the relative frequency of the respective p-values. The difference between the primary and secondary p-value sample sizes is due to the fact that not all clinical trials contain secondary study endpoints.

Our first hypothesis states that primary and secondary outcomes have inhomogeneous interval frequencies. We therefore test the stochastic independence of the observed p-curves by performing a chi-squared test on the frequency of p-value intervals and their position as primary or secondary outcomes. The results of this test led us to reject the null of statisti-

cal homogeneity between primary and secondary p-curves [DF=9 — TS=33. 0689— Prob. = 0.000]. The next hypothesis tackles anomalies in the p-curve. We assume that there is evidence of p-hacking in clinical trials if p-values are more likely to occur at the five percent level than at the preceding interval. Therefore, we conduct binomial tests for uniform distribution of those intervals for primary and secondary p-values and small and large intervals.

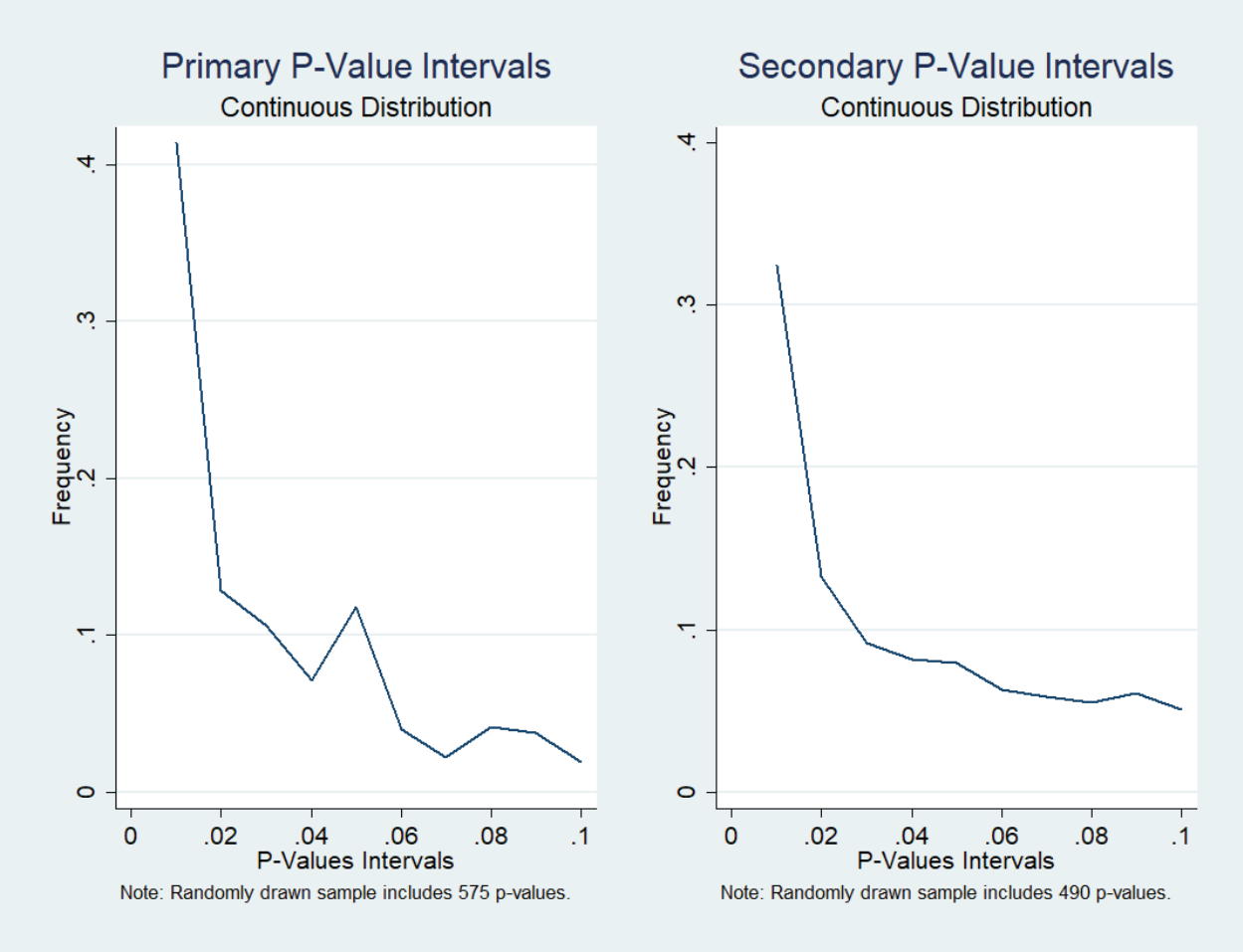

Figure 2: The p-curve is an observation of the frequency distribution of p-values from primary and secondary study endpoints up to the ten percent significance level. The abscissa displays the p-value intervals while the ordinate depicts the relative frequency of the respective interval. The difference between the primary and secondary p-value sample sizes is due to the fact that not all clinical trials contain secondary study endpoints.

As such, for the binomial test on the small primary intervals, we reject the null hypothesis [N=80 — Obs. p=.7375 — Pr (k 59) = 0.000]. For small secondary intervals, the null hypothesis of uniform distribution could not be rejected [N=116 — Obs. p=.57759 — Pr

(k 67) = 0.057]. The tests on the large intervals confirm this pattern. We could reject the null hypothesis on the large primary intervals [N=128 — Obs. p=.57812— Pr (k 74) = 0.046]. For the large secondary intervals, the null of uniform distribution could clearly not be rejected [N=175 — Obs. p=.48 — Pr (k 84) = 0.727].

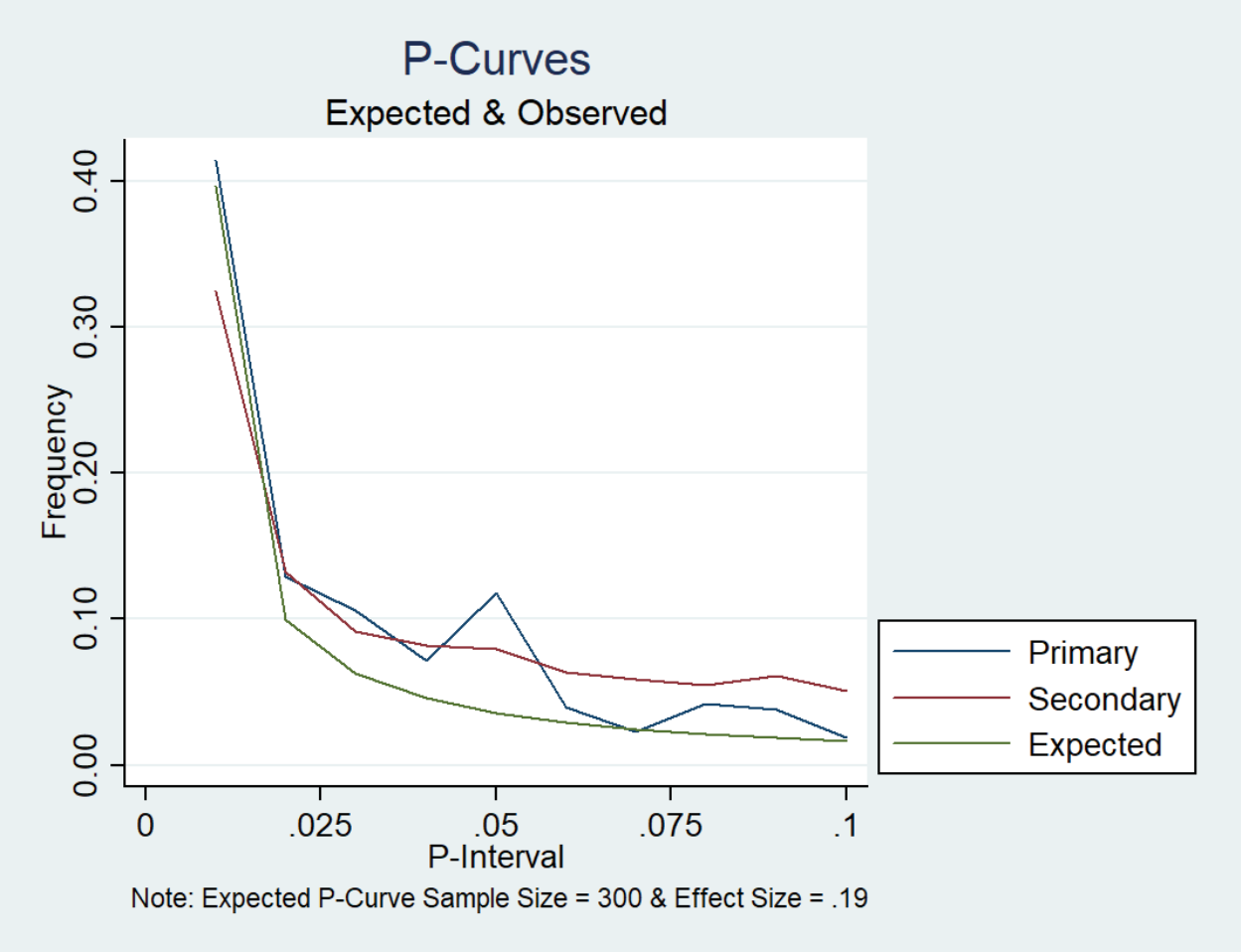

Figure 3: Primary, secondary, and expected p-curve up to the ten percent significance level. The abscissa displays the p-intervals while the ordinate depicts the relative frequency of the respective p-interval.

The findings reveal that the occurrences of primary p-values are more likely at the five percent level interval than at the preceding level. Figure 3 allows a direct comparison of the expected and observed p-curves. The expected p-curve is depicted by the green line and is based on a sample of 300 participants, considering an average effect size of .19 in clinical trials (Fukunaga & Kusama, 2014). The primary p-curve is represented by the blue line,

and the secondary p-curve is illustrated by the brown line. While the expected p-curve exhibits the lowest frequencies from the second interval on, both observed p-curves feature higher values virtually throughout all intervals with the exception of the first. Compared to the expected distribution, the secondary p-curve clearly features the ideal-typical shape of a monotonously decreasing function. The sole remaining distinctive feature is a flatter asymptotical approach to zero. Contrary to this, Figure 3 makes the anomaly of the primary p-curve even more obvious. All graphs and tests lead to a pronounced peak at the crucial five percent level of the primary p-curve.

4 Conclusion

In the current study, graphical examinations of the observed p-curves revealed strong anomalies in the distribution of primary p-values. The chi-squared test confirmed the statistical independence of primary and secondary p-curves. Further, by performing binomial tests, we proved significance for increasing frequencies of primary p-values left to the five percent level. Our findings indicate that there is a clear pattern in the data: Uniform distribution of the p-value data of the examined intervals is more presumable for the secondary results than it is for the primary results. The binomial tests highlight an anomaly in the shape of the primary p-curve, while the secondary p-curve corresponds to the ideal-typical shape of the expected p-curve under the assumption of evidential values.

We interpret the increasing frequencies of primary p-values left to the five percent level as evidence of p-hacking in clinical trials since primary outcomes are more decisive for the drug approval contrary to secondary outcomes, which are mainly tested due to their informative character (D'Agostino, 2000; Frantz, 2004). It is reasonable to assume that, despite the importance of ethical standards and statistical validity in clinical research, p-hacking is not merely hypothetical. We also would argue that researchers ambition of data-dredging is

limited (Simonsohn et al., 2014). P-hacking is only necessary until the p-value is less than or equal to the .05 level. This perfectly explains the peak at the five percent level even if we do not consider all inaccurately reported p-values in our analysis.

Regulations and compliance monitoring have major direct effects on the rejection rates and the time-to-market of newly developed drugs (Eichler et al., 2008). The more restrictive regulations are, the higher the probability that effective drugs will be rejected; vice versa, the less restrictive regulations are, the higher the probability ineffective drugs will be approved (Eichler et al., 2008). P-hacking might be no trivial offense but regulatory authorities, in general, are operating on a thin line between contrary objectives in respect to drug approval processes.

The p-value indicates the likelihood of a result occurring by chance alone. The null hypothesis significance testing is an arbitrary statistical construction that does not imply higher probability of a true hypothesis and indicates only how incompatible the data are with hypothesis (Wasserstein & Lazar, 2016). In the light of necessity of reproducibility, especially in medical science, the p-value is an imprecise tool for inference statistics since it is strongly influenced by sampling variability (Cumming, 2008).

Long-run progress in science needs corroboration. The short-run interests of publishing, funding, and research institutions are often in direct conflict with those long-run interests. A sound understanding hand in hand with proper application of statistics and robust experimental designs can help to cure the problem of data dredging and simultaneously enhance ethical standards. It is about time that incentives are established and institutions make stakeholders recognize that long-run needs are in their own best interests.

5 References

Al-Marzouki, S., Evans, S., Marshall, T., & Roberts, I. (2005). Are these data real? Statistical methods for the detection of data fabrication in clinical trials. *Bmj*, 331(7511), 267-270.

Altman, N., & Krzywinski, M. (2017). Points of significance: P values and the search for significance. *Nature Methods*, 14(1), 3-4.

Bradshaw, M.; Guarino, R. A., & Guarino, R. (Eds.). (2016). *New drug approval process*. CRC Press.

Buyse, M., George, S. L., Evans, S., Geller, N. L., Ranstam, J., Scherrer, B., ... & Colton, T. (1999). The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. *Statistics in medicine*, 18(24), 3435-3451.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., ... & Heikensten, E. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280), 1433-1436.

Collins, F. S., & Tabak, L. A. (2014). NIH plans to enhance reproducibility. *Nature*, 505(7485), 612.

Council for International Organizations of Medical Sciences. (2002). International ethical guidelines for biomedical research involving human subjects. *Bulletin of medical ethics*, (182), 17.

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, 3(4), 286-300.

D'Agostino, R. B. (2000). Controlling alpha in a clinical trial: the case for secondary endpoints. *Statistics in Medicine*, 19(6), 763-766.

DiMasi, J. A., Hansen, R. W., & Grabowski, H. G. (2003). The price of innovation: new estimates of drug development costs. *Journal of health economics*, 22(2), 151-185.

Eichler, H. G., Pignatti, F., Flamion, B., Leufkens, H., & Breckenridge, A. (2008). Balancing early market access to new drugs with the need for benefit/risk data: a mounting dilemma. *Nature Reviews Drug Discovery*, 7(10), 818-826.

Emanuel, E. J., Wendler, D., & Grady, C. (2000). What makes clinical research ethical?. *Jama*, 283(20), 2701-2711.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PloS one*, 4(5), e5738.

Farland, L. V., Correia, K. F., Wise, L. A., Williams, P. L., Ginsburg, E. S., & Missmer, S.A. (2016). P-values and reproductive health: what can clinical researchers learn from the American Statistical Association?.

Frantz, S. (2004). Lessons learnt from Genasense's failure. *Nature Reviews Drug Discovery*, 3(7), 542-542.

Fukunaga, S., Kusama, M., & Ono, S. (2014). The effect size, study design, and development experience in commercially sponsored studies for new drug applications in approved drugs. *SpringerPlus*, 3(1), 740.

George, S. L., & Buyse, M. (2015). Data fraud in clinical trials. *Clinical investigation*, 5(2), 161.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS biology*, 13(3), e1002106.

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

Ioannidis, J. P., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical trials*, 4(3), 245-253.

Ioannidis, J. P., Allison, D. B., Ball, C. A., Coulibaly, I., Cui, X., Culhane, A. C., ... & Mangion, J. (2009). Repeatability of published microarray gene expression analyses. *Nature genetics*, 41(2), 149-155.

Jager, L. R., & Leek, J. T. (2013). An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*, 15(1), 1-12.

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *The Quarterly Journal of Experimental Psychology*, 65(11), 2271-2279.

Meinert, C. L. (2012). *ClinicalTrials: Design, Conduct and Analysis*. Oxford University Press.

Miller, J. E., Korn, D., & Ross, J. S. (2015). Clinical trial registration, reporting, publication and FDAAA compliance: a cross-sectional analysis and ranking of new drugs approved by the FDA in 2012. *BMJ open*, 5(11), e009758.

Mullard, A. (2015). 2014 FDA drug approvals. *Nature Reviews Drug Discovery*, 14, 7781.

National Institutes of Health. (2016). NIH Policy on dissemination of NIH-funded clinical trial information. *Fed Regist*, 81(183).

Newman, M. E. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary physics*, 46(5), 323-351.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.

Powell-Smith, A., & Goldacre, B. (2016). The TrialsTracker: automated ongoing monitoring of failure to share clinical trial results by all major companies and research institutions. *F1000Research*, 5.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets?. *Nature Reviews Drug Discovery*, 10(9), 712-712.

Ranstam, J., Buyse, M., George, S. L., Evans, S., Geller, N. L., Scherrer, B., ... & Colton, T. (2000). Fraud in medical research: an international survey of biostatisticians. *Controlled clinical trials*, 21(5), 415-427.

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

Saito, H., & Gill, C. J. (2014). How frequently do the results from completed US clinical trials enter the public domain? A statistical analysis of the ClinicalTrials.gov database. *PLoS One*, 9(7), e101826.

Scannell, J. W., Blanckley, A., Boldon, H., & Warrington, B. (2012). Diagnosing the decline in pharmaceutical R&D efficiency. *Nature Reviews Drug Discovery*, 11(3), 191-200.

Seife, C. (2015). Research misconduct identified by the US Food and Drug Administration: out of sight, out of mind, out of the peer-reviewed literature. *JAMA internal medicine*, 175(4), 567-577.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11), 1359-1366.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology*: General, 143(2), 534.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose.

Weir, C., & Murray, G. (2011). Fraud in clinical trials. *Significance*, 8(4), 164-168.

Otto von Guericke University Magdeburg
Faculty of Economics and Management
P.O. Box 4120 | 39016 Magdeburg | Germany

Tel.: +49 (0) 3 91 / 67-1 85 84
Fax: +49 (0) 3 91 / 67-1 21 20

www.fww.ovgu.de/femm

ISSN 1615-4274