relationships[6] by enabling systematic analysis of large-scale phenotype data.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nmeth.3477).

ACKNOWLEDGMENTS
The US National Institutes of Health (NIH) funded Phenoview (U54 HG006370). We thank the phenotyping centers for collecting the data, and EMBL-EBI for hosting the data archive.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Gagarine Yaikhom, Hugh Morgan, Duncan Sneddon, Ahmad Retha, Julian Atienza-Herrero, Andrew Blake, James Brown, Armida Di Fenza, Tanja Fiegel, Neil Horner, Natalie Ring, Luis Santos, Henrik Westerberg, Steve D M Brown & Ann-Marie Mallon

MRC Harwell, Medical Research Council, Harwell, UK.
e-mail: g.yaikhom@har.mrc.ac.uk

1. Brown, S.D.M. & Moore, M.W. *Dis. Model. Mech.* **5**, 289–292 (2012).
2. Koscielny, G. *et al. Nucleic Acids Res.* **42**, D802–D809 (2014).
3. Boyle, J. *Nature* **499**, 7 (2013).
4. Frankel, F. & Reid, R. *Nature* **455**, 30 (2008).
5. Karp, N.A., Melvin, D., Sanger Mouse Genetics Project & Mott, R.F. *PLoS ONE* **7**, e52410 (2012).
6. Dowell, R.D. *et al. Science* **328**, 469 (2010).

Clarifying the terminology that describes scientific reproducibility

To the Editor: There has recently been a growing interest in discussions of reproducible/repeatable scientific research[1,2]. The scientific press appears to be witnessing a confusion of terms: reproducibility, repeatability and replicability are referred to with different and sometimes conflicting meanings, both between and within fields. We suggest that these terms can be clarified by considering the intended generalization of the study at hand.

In industrial systems, for example, reproducibility and repeatability are used in the context of 'gauge repeatability and reproducibility' (GR&R) testing for evaluating measurement error of equipment. In these experiments, several testers are asked to retest a set of items. Differences between testers under different conditions are used to estimate reproducibility, whereas repeat evaluations under identical conditions are used for estimating repeatability[3]. In contrast, the term replicability is used in genome-wide association studies to describe a repetition of a study by the same lab or researchers but with a different technology or a different data set (typically a follow-up subpopulation but possibly a different human population)[4], whereas in GR&R, such a case would be called repeatability. In machine learning and computational mathematics, experiments are used to evaluate algorithms. The common terms in these fields are reproducibility and replicability, but different researchers have different definitions[5]. One distinction is whether the exact numerical results are recreated—for instance, by rerunning the code (repeatability)—versus whether the overall result can be rederived (replicability). Finally, in preclinical studies, reproducibility often relates to recreating the same numbers by different labs, whereas in GR&R and machine learning, the same term is used to describe changing experimental conditions beyond the researchers or lab. We see that the same terms are used with different meanings in different contexts. Our goal here is to provide conceptual clarification to this situation.

Reproducibility, replicability and repeatability are defined by which experimental conditions are changed versus which are kept constant, but definitions vary across areas. We suggest that these terms can be clarified by considering the intended generalization of the study. Generalization is a key concept in the information quality, or InfoQ, framework that we proposed in the context of applied research[6] and is, in some form, part of the goal of every scientific study. Statistical generalizability refers to inferring from a sample to a target population. Statistical analyses performed in scientific studies are typically aimed at achieving statistical generalizability. Scientific generalizability, on the other hand, refers to applying a model based on a particular target population to other populations. Reproducibility, repeatability and replicability are aimed at assuring generalizability, but the generalizability is typically of different types.

To illustrate this, we consider again the GR&R case. The goal behind GR&R repeatability is to assess measurement error of a specific device or technology for future use (statistical generalization). Therefore, test conditions are kept constant, but multiple testers are employed in rerunning tests of specific test items. Poor repeatability indicates needed improvement of the measurement technology. In contrast, the goal of GR&R reproducibility is generalizing to future use under different testing conditions such as different lab technicians or test environments (scientific generalization), and therefore both test conditions and testers are varied. Poor reproducibility calls for considering the overall measurement process, including operating procedures and provided training.

As an example from biological studies, we consider the recent criticism of standardization in animal behavior experiments[7]. The authors show that, in contrast to standardization being beneficial, introducing systematic variation of experimental conditions (which they call "heterogenization") may attenuate spurious results and improve reproducibility[8]. Considering this from the standpoint of generalization clarifies the issue. Standardized animal behavior experiments are differently generalizable than experiments with induced systematic variation of experimental conditions. In particular, standardization intends statistical generalization, whereas heterogenization intends scientific generalization.

In summary, although terminology can remain domain specific, we propose that researchers should clearly state the intended generalization of their study. Such an approach will clarify the implications of a study within and across fields.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Ron S Kenett[1,2] & Galit Shmueli[3]

[1]University of Turin, Turin, Italy. [2]The KPA Group, Raanana, Israel. [3]Institute of Service Science, National Tsing Hua University, Hsinchu, Taiwan.
e-mail: ron@kpa-group.com

1. McNutt, M. *Science* **343**, 229 (2014).
2. Banks, D. *Stat. Politics Policy* **2**, doi:10.2202/2151-7509.1023 (2011).
3. Kenett, R.S., Zacks, S. & Amberti, D. *Modern Industrial Statistics: With Applications in R, MINITAB and JMP* 2nd edn. (Wiley, 2014).
4. Ionnides, J.P. *et al. Nat. Genet.* **41**, 149–155 (2009).
5. Drummond, C., Japkowicz, N., Klement, W. & Macskassy, S.A. in *Proc. 26th. Int. Conf. Mach. Learn.* doi:10.1145/1553374.1553546 (ACM, 2009).
6. Kenett, R.S. & Shmueli, G. *J. R. Stat. Soc. Ser. A Stat. Soc.* **177**, 3–38 (2014).
7. Richter, S.H., Garner, J.P. & Würbel, H. *Nat. Methods* **6**, 257–261 (2009).
8. Richter, S.H., Garner, J.P., Auer, C., Kunert, J. & Würbel, H. *Nat. Methods* **7**, 167–168 (2010).