The impact of shared data in neuroimaging: the case of OpenfMRI.org

Krzysztof J. Gorgolewski[1], Kelsey Wheeler[2], Yaroslav O. Halchenko[2], Jean-Baptiste Poline[3], Russell A. Poldrack[1]

1. Department of Psychology, Stanford University, Stanford, CA, USA; 2. Department of Psychology and Brain Sciences, Dartmouth College, Hanover, NH, USA; 3. Helen Wills Neuroscience Institute, University of California at Berkeley, CA, USA

chrisgor@stanford.edu

poldracklab.stanford.edu

Stanford University

Introduction

Data sharing in neuroimaging has been advocated for almost 15 years [1–3]. Sharing data can in principle lead to more transparency, fewer methodological mistakes, new findings, as well as savings in data acquisition costs. On the other side some skeptics claimed in the past that “fMRI data per se are very context sensitive” which ”somewhat subverts the raison d’être for sharing” [4]. Van Horn and Gazzaniga in their summary of fMRIDC highlight five successful studies reusing data deposited in their database [5]. However, there has been no systematic survey of reuses of publicly available neuroimaging data. In this study we focus on the OpenfMRI database [6] and the reuses of datasets deposited in it.

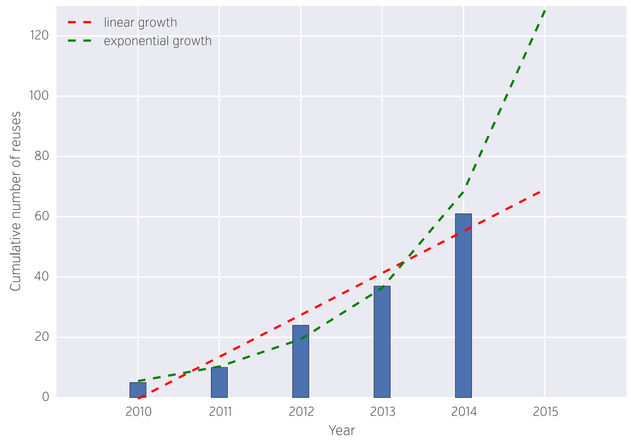
Methods

- we used Scopus and Google to find all scientific products citing the oldest paper associated with each dataset
- 24 curated OpenfMRI datasets were included in the analysis
- we also reached out to dataset authors asking about any reuses known to them
- we excluded papers published before the dataset was publicly available
- papers whose authors overlap with the reused dataset were also excluded
- the remaining 2421 papers were evaluated for data reuse
- each reuse was evaluated in terms of topic, number of datasets used and date of publication

We also attempted to estimate how much money was saved by reusing data. To do this we calculated hypothetical cost of acquiring new data for all studies reusing datasets from OpenfMRI.org. To simplify calculation we have used a conservative fixed estimate of $600 per subject accounting for the cost of designing, piloting and running the study, as well as compensation for participants.

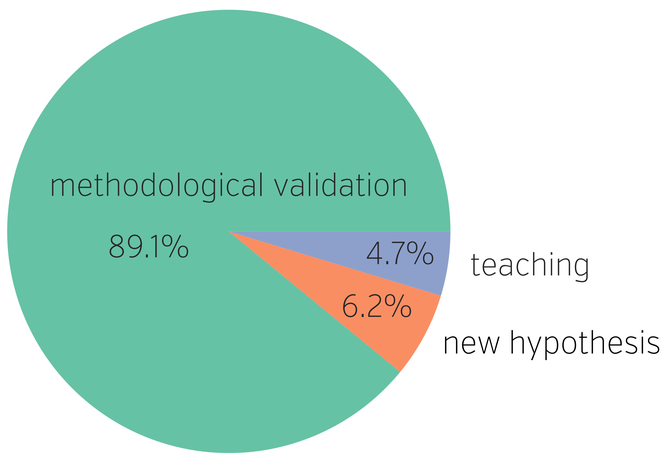
Results

1



Cumulative distribution of dataset reuses in the past five years with linear and exponential growth projections. A reuse is defined as a paper, preprint or tutorial using at least one OpenfMRI dataset.

2



Distribution of topics represented by discovered reuses. Data deposited in OpenfMRI has been mostly used for showcasing and validating methodological advancements. A small percentage of cases used the data for supporting new discoveries or teaching.

Results cont.

63 number of papers and courses using OpenfMRI data

60% of papers used more than one dataset

39 number of reuses of the most popular dataset [7]

$878,400

how much it would cost to perform those studies without shared data

References

1. Van Horn JD et al. (2001) The Functional Magnetic Resonance Imaging Data Center (fMRIDC): the challenges and rewards of large-scale databasing of neuroimaging studies. Philos Trans R Soc Lond B Biol Sci 356: 1323–1339.

2. Poline J-B et al. (2012) Data sharing in neuroimaging research. Front Neuroinform 6: 9.

3. Poldrack RA, Gorgolewski KJ (2014) Making big data open: data sharing in neuroimaging. Nat Neurosci 17: 1510–1517.

4. Friston KJ (2009) Modalities, modes, and models in functional neuroimaging. Science 326: 399–403.

5. Van Horn JD, Gazzaniga MS (2013) Why share data? Lessons learned from the fMRIDC. Neuroimage 82: 677–682.

6. Poldrack RA et al. (2013) Toward open sharing of task-based fMRI data: the OpenfMRI project. Front Neuroinform 7: 1–12.

7. Haxby JV, et al. (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293: 2425–2430.