OPEN ACCESS

CrossMark

Research Idea

A Million Brains in the Cloud

Arno Klein[‡], Satrajit S. Ghosh[§]

‡ Sage Bionetworks, Seattle, United States of America
§ MIT, Cambridge, United States of America

Corresponding author: Arno Klein (arno@binarybottle.com)

Reviewable   v1

Abstract

Background

There is a great need for more publicly available human brain image data sets to establish baseline variation in structure, function, and connectivity in a given population, and to take advantage of powerful new data mining tools that require large sample sizes.

New information

In the following, we propose to engage citizens to take ownership of the images of their own bodies, and to provide them with a means of volunteering these images for research and collective analysis. We propose to build a platform with an intuitive interface to make it extremely easy for people to direct their brain images to a central repository, with the goal of accruing brain images from an unprecedented one million individuals.

Keywords

brain imaging, data mining

Overview and background

Human brain images are being generated at an ever increasing scale and rate. Yet the largest publicly available data set contains only a few thousand participants, and it is standard for a brain imaging study to enlist mere dozens of participants. These focused studies attempt to control for variation in a given population in order to measure differences in signal due to a factor of interest. This narrow focus is precisely the reason why we know so little about the variation in structure, function, and connectivity in a sample of people, let alone in the population at large. Without establishing baseline variation, it is folly to think such miniscule numbers of brains will give us an understanding of the differences among signals extracted from brains. Our most powerful analytical tools for mining high-dimensional data are helpless with such small sample sizes.

Amassing large numbers of brain images is absolutely critical for establishing variation among brains, for recategorizing or distinguishing subtypes of disorders in a clinical population, for generating data sets of otherwise underrepresented subpopulations (demographic minorities, rare disorders, etc.), and for reevaluating null results in prior studies. It will take significant sociological change to provoke hospitals to share the data they acquire even with their patients' consent. In the meantime we think that it might be more fruitful to engage citizens to take ownership of the images of their own bodies, and to provide them with a means of volunteering these images for research and collective analysis. We propose to build a platform with an intuitive interface to make it extremely easy for people to direct their brain images to a central repository, with the goal of accruing brain images from an unprecedented one million individuals. Not only would this number of images help to address the challenges above, it would draw attention to variation itself as a subject of study and help establish baseline variation as the foundation for future research and clinical practice. We would work with clinical centers and research laboratories to allow their patients and participants to upload their data at the time of collection.

There are numerous potential incentives we could employ to engage these individuals, such as offering them the chance to own and share their electronic medical records, to receive free clinical feedback based on these images, to support or be an active participant in scientific endeavors, or even to receive a token based on their involvement, such as a 3-D printed sculpture of their own brain.

Ultimately, we believe the future of medical data will involve people participating on the same data-exchange playing field as their doctors and biomedical researchers. This is part of a Personal Data Store vision, where individuals would self-report data and determine what information is gathered, who may have access to it, and how it may be used. In addition to brain scans acquired in a research or clinical setting, such data would be acquired via wearable devices or smart phone apps. From a data analysis perspective, these remote sensors would enable massive collection of survey, biometric, and behavioral data. From an individual's perspective, such personal and population-wide data collection can assist in coaching, counseling, clinical monitoring, and biofeedback training. For individuals, this would enable a level of assistance and intervention not possible without a

round-the-clock personal coach, counselor, and caretaker, alerting the individual, software, or possibly public health officials of sudden changes, troubling trends, or predictors of upcoming problems.