Published in final edited form as:

Psychol Aging. 2022 February ; 37(1): 141–147. doi:10.1037/pag0000605.

Data Overuse in Aging Research: Emerging Issues and Potential Solutions

Daniel K. Mroczek[1,2], **Sara J. Weston**[3], **Eileen K. Graham**[1], **Emily C. Willroth**[1]

[1]Northwestern University, Feinberg School of Medicine, Dept. of Medical Social Sciences

[2]Northwestern University, Weinberg College of Arts & Sciences, Dept. of Psychology

[3]University of Oregon, Dept. of Psychology

Abstract

Aging and lifespan development researchers have been fortunate to have public access to many longitudinal datasets. These data are valuable and see high utilization, yet this has a considerable downside. Many of these are heavily overused. Overuse of publicly-available datasets creates dependency among published research papers giving the false impression of independent contributions to knowledge by reporting the same associations over multiple papers. This is a potentially serious problem in the aging literature given the high use of a relatively small number of well-known studies. Any irregularities or sampling biases in this relatively small number of samples has outsize influence on perceived answers to key aging questions. We detail this problem, focusing on issues of dependency among studies, sampling bias and overfitting, and contradictory estimates of the same effect from the same data in independent publications. We provide solutions, including greater use of data sharing, pre-registrations, holdout samples, split-sample cross-validation, and coordinated analysis. We argue these valuable datasets are public resources that are being diminished by overuse, with parallels in environmental science. Taking a conservation perspective, we hold that these practices (pre-registration, holdout samples) can preserve data resources for future generations of researchers.

Keywords

Data Overuse; Coordinated Analysis; Open Science; Replicability; Adult Development and Aging

The “tragedy of the commons” refers to overuse of shared resources to the point of depletion, often unrecoverable depletion. Social and biomedical scientists, especially aging and lifespan development researchers, are fortunate to have access to many shared resources in the form of publicly-accessible datasets that are both large-scale and long-term. Can such resources be depleted by overuse? Perhaps not in the same way as hunting the red

Correspondence concerning this article should be addressed to Daniel K. Mroczek, Northwestern University, Dept. of Medical Social Sciences, Feinberg School of Medicine, 625 N. Michigan Ave, 22nd Floor, Chicago IL 60611. daniel.mroczek@northwestern.edu.

Author Contributions. All authors contributed to the conceptualization and preparation of this manuscript. Author Mroczek prepared the initial draft and authors Weston, Graham, and Willroth added further material along with edits and revisions. All 4 authors contributed to final edits and revisions.

Prior Presentation: Portions of this manuscript were presented at the Gerontological Society of America conference, November 2019.

elk to extinction in 18th century North America or overfishing swordfish in the Atlantic to the point of population collapse. Rather, the overuse of public data leads to a deficiency in the resulting scientific literature. For example, overuse creates dependencies among published research papers (Thompson et al., 2020) giving a false impression of independent contributions to knowledge by reporting of the same associations in multiple papers. This depletes the value of the scientific literature, thereby reiterating the tragedy of the commons in a manner not quite seen before.

In this paper, we discuss the problem of scientific data overuse and possible solutions. Scientific data overuse is the repeated and excessive use of a single dataset, giving rise to many published findings that are not independent of one another. We first take stock of common data resources, focusing on longitudinal lifespan-oriented studies, and why they are considered valuable. We then turn to examples of the consequences of using a single dataset for multiple studies, thus illustrating the problem in a concrete manner. Next, we propose two types of solutions: one depends on scientists self-regulating in an honor system approach and one is more heavy-handed and involves potential regulatory bodies. We conclude on an optimistic note, by discussing that the credibility movement has demonstrated how scientists can change practices rapidly and for the better. We take a conservation perspective and argue that by using some of the proposed solutions (e.g., preregistration, holdout samples, coordinated analysis) we can preserve these data resources for future generations of researchers. However, first we define what we mean by data overuse. Scientific data overuse is the *repeated and excessive use of a single dataset, giving rise to many published findings that are not independent of one another*.

Public Data Are Public for a Reason

Data are generally placed in the public domain because they are valuable. This accessibility happens through voluntary action of the study leaders but also via mandate by funding agencies. By “valuable” we mean that data collection is resource intensive. Valuable datasets may have multiple measurement occasions, be comprised of large and diverse samples and subsamples, or possess an extensive number of variables (hundreds, or even thousands in some cases) as well as a wide variety of variable types (informant and self-reports; biological as well as psychological constructs; linked data such as O*Net or Medicare linkages). This is usually a lot more than any single researcher can collect working alone. For these reasons, many datasets enjoy enhanced valuation in the eyes of scientists. Datasets possessing such characteristics – for example Health and Retirement Study (HRS), Religious Orders Study (ROS), Midlife in the U.S. Study (MIDUS), Midlife in Japan Study (MIDJA), the Chinese Family Panel Study, the Korean Study of Women & Families, the German Socio-Economic Panel (GSOEP), and UK Biobank – are made public precisely because of these valuable dimensions.

In many cases, the size, quality, and interdisciplinary nature of these datasets are beyond what a single scientist could ever manage to assemble while working in the typically small-scale and underfunded social science laboratory. Major longitudinal datasets featuring many different types of data require large interdisciplinary teams and significant monetary resources. This creates yet another pressure for making data publicly available, as taxpayer money often funds these resources. The long-term longitudinal nature of such data also requires intergenerational cooperation between younger and older researchers, as many studies can span 50 or more years, thus exceeding the career spans of individual scientists, a point we have argued elsewhere (Mroczek et al., 2011).

Given the value of these types of datasets to scientists and taxpayers, funding bodies have instituted data sharing policies, creating additional pressure for making data public. Starting in the 1990s, various agencies – but in particular the U.S. National Institutes of Health – began mandating that more expensively-collected datasets be deposited in a public repository such as ICPSR (Inter-university Consortium of Political & Social Research) not long after collection. Eventually this was extended to less-expensively collected data as well. By the 2010s, researchers had many public datasets to choose from, although the ones with heaviest use tended to have more of the desirable qualities described above (e.g., more waves, bigger samples, greater variety of variables).

These datasets represent a valuable public commodity. Public data resources are a common good quite similar to shared natural resources (forests, lakes, rivers, oceans) or built resources (public transportation, roadways and sidewalks, public buildings). However, like these shared natural and built resources, problems usually arise when resources are overused. Such is the case with data resources.

The Problem of Data Overuse

Why is overuse of data a problem? We focus on three problems that either arise due to overuse or are exacerbated by it. These are: 1) dependency among published findings, 2) amplification of sample peculiarities (i.e., sampling error), and 3) contradiction. Some of these problems are unique to data overuse (e.g., contradictory published findings from the same datasets) but others are more general problems (e.g., sample peculiarities). In the latter, we maintain that data overuse exacerbates and amplifies long-standing problems. Aging or lifespan samples that are unrepresentative biased in some way (e.g., tend to be well-educated or underrepresent certain disadvantaged groups) may inflate confidence in findings due to the excessive use of a small group of datasets. One study is only one study, but data overuse can create the appearance that one study is five studies; this gives the impression that a particular finding has been replicated, when it fact, it has only been reproduced (Condon, Graham, & Mroczek, 2018). This falsely increases confidence in a finding that hasn't been independently replicated. In addition, our confidence in findings can be inflated in other ways by data overuse. If each dataset had only been used once, would the magnitude or reach of a given problem be the same? For example, the problem of sampling error and overfitting (discussed below) may still be present, but the magnitude of the problem may not be as great if data overuse had not amplified or exacerbated it. We describe each of the three problems in the next sections.

Dependency

The progression from initial report to accepted scientific finding typically requires multiple studies that together establish replicability and may also probe the limits of generalizability. Such a body of studies is the hallmark of cumulative science. Yet cumulative science assumes that most of these studies are independent of one another. This is compromised when many of the findings are from the same small group of datasets.

Just as most statistical tests assume independent individual observations within a given study, similarly we assume independent studies within a body of findings when evaluating that literature. However, if a particular dataset appears frequently, there is some degree of dependency among published studies. In a single investigation, non-independent observations can raise the chance of a false positive; in a summary of a field, non-independent studies also lead to a higher false positive rate (Thompson et al., 2020).

Some have proposed remedies, mostly focused on adjusting the alpha level across multiple studies using the same dataset. This is often referred to as “sequential” correction of alpha, in contrast to “simultaneous” correction for multiple tests within a single investigation or manuscript (Foster & Stine, 2008). In sequential correction, the goal is to gain some degree of controllability over the false positive rate when dealing with many investigations that have accrued over many years that use the same data. This solution is fallible because it requires keeping track of all studies that use the same dataset. This can be difficult with widely used data that may be used hundreds or thousands of times. As an example, the topic “affect” has accrued 307 publications in the MIDUS study over the past 25 years. Thompson et al. (2020), focusing on public neuroscience data and labelling the problem “data decay,” proposed a different kind remedy known as alpha debt. This involves prospectively extending the adjustment of alpha criteria via Bonferroni correction to all publications using the same dataset. The first study may use alpha of .05, but the second uses .025, and so on. This may be done prospectively as manuscripts accrue in real time.

An issue with sequential and alpha debt correction methods is that they focus on alpha level at a time when many areas of science are moving away from null hypothesis significance testing (NHST) and its reliance on *p*-levels. Although these techniques recognize the dependency problem and its propensity to raise the false positive rate, they remain tethered to the NHST framework. In the solution section we suggest other ways of dealing with the dependency issue.

One manifestation of the dependency problem that is exacerbated by data overuse is multiple entries of the same association from the same dataset into meta-analyses. We point out that although this can happen even if a given dataset is not particularly overused and may not be a problem generated solely by overuse. It is, nonetheless, an issue that is amplified or made worse by excessive use of certain datasets. It is also the case that a careful meta-analyst should be able to weed redundant studies. Thus to some degree, it is the responsibility of the individual meta-analyst to keep this problem at bay.

We consider multiple entries a salient problem because meta-analyses are often taken as authoritative. Yet if they contain redundancies, then the meta-analytic summary statistics are compromised, and the overall conclusions of the meta-analysis are therefore biased. This can mislead scientists more than a single paper would. As an example, in a review of personality traits and dementia risk (Low et al., 2013), there were four entries of the Memory and Aging Project (MAP) and two entries of the Religious Orders Study (ROS). This likely inflated the summary statistics (hazard ratios) by basing them on an overall meta-analytic sample size that was far higher than was actually the case. A solution would have been to count the ROS and MAP only once, even though each had multiple papers in the literature on the effects of interest. This would have decreased the total number of studies entered into the meta-analysis but also would have provided a more realistic effect size estimate. In some of these instances given MAP studies were published several years apart, thereby allowing new dementia cases to accrue. New journal articles based on updated data are reasonable. However, there should be some correction for this when meta-analyses are performed. After all, they represent the same respondents, and the baseline personality measures are the same.

Sample Peculiarities and Overfitting

When samples are repeatedly used to answer a considerable proportion of the research questions in a given area, a danger arises that the answers will be biased. It is the nature of statistics that every sample from a population contains sampling error. With repeated, independent draws from the population, sampling errors across samples are expected to average to 0. However, if one sample contributes unequally, its sampling error will remain unbalanced, thus pushing a meta-analytic effect away from the true population parameter.

For example, much of what we know about how early childhood events interact with genetic factors to predict adulthood psychopathology come from the Dunedin sample (Caspi et al., 2003; Moffitt et al., 2010). Thus, any irregularities due to sampling error in this particular sample could have outsize influence on answers to key research questions. The Dunedin study is one that has high valuation, as defined earlier. It has data from early childhood through middle age and rich biological, as well as psychological, data at many waves. That said, it is one sample and has sampling error, and any single sample may be unrepresentative or skewed in ways that are hard to identify in advance or even many years into data collection. This may lead to incorrect answers to research questions. For example, the Dunedin study was the first to report a particular gene-by- environment interaction – the 5-HTT by stressful life events – on several later indicators of psychopathology (Caspi et al., 2003). This was considered a very important finding at the time and in the years that followed, and there were many attempts to replicate with other samples. Most of these replication attempts failed to confirm the original finding (Risch et al., 2009; Munafò et al., 2009).

As there is no reason to suspect computational or data management errors in the original study, the Dunedin result was the correct result for that sample and in this sense was reportable as a legitimate finding, made in good faith. Yet the finding may have been biased by the peculiarities of the sample that was used and therefore was *overfit* to that particular sample (Babyak, 2004; Yarkoni & Westfall 2017). Overfitting occurs when the results of modeling are too closely tied to the dataset being used and is a common problem in psychology, engineering, investment and financial modeling, and other areas. Among statisticians, overfitting often specifically refers to a model too complex for the data, such as one with too many polynomial terms or a piecewise model with too many spline parameters. However, the definition may be generalized to refer to any circumstance in which findings become too tailored to the quirks and random noise of a given sample.

This points to a larger issue that is relevant to data overuse, and this is the distinction between replication and generalizability (Condon, Graham & Mroczek, 2018). There is a need to not overgeneralize findings to populations not sampled. A lack of replication may not reflect replicability at all but rather a generalizability problem stemming from the types of populations (often middle class and from Western countries; Henrich, Heine, & Norenzayan, 2010) that often make up overused samples.

Of course, sample peculiarity, overfitting and lack of generalizability are problems that plague research in general and are not necessarily an issue unique to data overuse. Some have argued that bias and generalizability issues are pervasive in all kinds of research (Yarkoni, in press). That said, these problems are particularly magnified by data overuse. Every sample has peculiarities, such as skewed distributions, non-representativeness, and selection biases of all kinds. That is simply the nature of sampling and data collection. No dataset will be perfectly representative of the human population. Replication in many independent samples carried out by independent groups of researchers confirms whether an initial finding holds up or was just a fluke of the first sample used. However, if that same sample is used over and over again, then those small peculiarities can inadvertently place inaccurate, misleading or non-generalizable findings into the scientific literature, leading researchers astray. Excessive use of a few well-known datasets also creates the appearance of replicability or generalizability for unreplicable or ungeneralizable findings, in turn inflating our confidence in those results. For all of the reasons discussed above, data overuse can impact generalizability. That said, authors using widely-used datasets should acknowledge these limits in a Constraints on Generality (COG) section within the discussion portion of each paper (Simons, Shoda, & Lindsay, 2017). An author could state that in using a given widely-used dataset, they are potentially capitalizing on biases that stem from the unique quirks and peculiarities of that sample, and this can place a constraint on generalizability. An author can go on to state that similar analysis using other samples are needed to better evaluate both the replicability and generalizability of the effect under investigation. It clarifies when an investigator expects their results to generalize, and this tempers the interpretation by readers that a particular study could describe all persons.

Many overused datasets are longitudinal and ongoing, meaning they are frequently being updated with new waves of data from the same participants. Additional measurement occasions increase the value of a dataset, as do mortality updates or linkages with administrative sources and existing records (census, birth registry, Medicare, Social Security; Ferrie et al., 2012). Updating and continuing longitudinal data collection can mitigate the overfitting issue to some extent but does not solve the fundamental underlying problem. Even if a scientific team obtains more data on participants, they are still based on the same people. Moreover, if replication studies combine early and newer waves, the earlier waves are still the same waves as the original finding.

Refreshing the sample with new cohorts can provide a stronger hedge against the overfitting and bias issue. Yet this is expensive and time consuming and very few studies have done so. The Seattle Longitudinal Study (SLS), the MIDUS, and the HRS are notable exceptions. That said, most of these refreshers are few and far between and most were not done until a decade or more after the initial sample data collection. In sum, new recruitment and refreshment can potentially provide an antidote against overfitting, but it does not solve the fundamental underlying problem.

Contradiction

Our third problem refers to contradictory findings in different papers using the same dataset that remain unaddressed or unidentified. While it is possible for any dataset to be used for different publications and show contradictory findings for the same research question, this issue becomes more likely as a given datasets is repeatedly used. Importantly, it is not the contradictory findings themselves that prevent researchers from making robust conclusions but rather the lack of understanding of why two analyses yield different results.

Often, when a given dataset is used again and again, the same research question will be tested in separate papers, often not citing one another. Due to researcher degrees of freedom (Wicherts et al, 2016; Simmons et al., 2011), different analyses will make use of varying sets of covariates, different measurement waves, and different statistical models (e.g., OLS regression vs. structural equation models). These choices can make a small effect disappear and reappear across different papers.

For example, two papers in the past decade reported a statistical test of the association between agreeableness and verbal fluency. Graham and Lachman (2012) reported a significant ($p < .05$) negative association, while Sutin and colleagues (2019) found no effect. Both papers used the MIDUS wave 2 data to test these associations using linear regression but included different sets of covariates in their models: both papers controlled for age, sex, and education, while Sutin et al. (2019), included race, and Graham et al. (2012) included self-reported health. This difference in just two covariates was enough to pull the confidence interval into the significant range for one, but not the other. This example highlights the inconsistent identification of covariates within sub-disciplines of psychology, as well as the limited use of cross-validation and sensitivity analyses in the field. More importantly, contradictory findings in the same data are valuable and provide opportunities for discussing norms and assumptions in a research domain. However, this discussion can only occur when researchers and readers are aware of the multiple use of data across studies

The previous finding of Graham et al. (2012) was not discussed (but was cited) in the more recent paper (Sutin et al. 2019), which highlights another problem: over-proliferation of scientific papers. Of course, this touches on the separate issue of academic incentives that sometimes value quantity over quality. Easy data access means many more published manuscripts than otherwise would be the case. Crowded literatures make it easy to miss papers with similar analyses to the ones a researcher is planning for a new manuscript. Confusion results when findings in different papers using the same dataset contradict one another and all of them make their way into the extant literature.

To some extent, examples such as the one we described above may be the result of journal page limits and reference caps. We suggest that caps on number of references are counterproductive. Scientific policies should encourage near-exhaustive literature reviews in advance of carrying out a research project. We revisit this point in the solutions section below.

We wish to point out that the above scenario is different from what statisticians call a sensitivity analysis (Chatterjee, 2009; Saltelli et al., 2004). In a rigorous investigation of the sensitivity of a given effect, different covariates are entered in careful stages. If an association is rendered spurious when controlling for a given confound, then we obtain an idea of when an effect is likely to hold in re-analyses of the same datasets in the future or in new samples. Of course, even a well-performed sensitivity analysis is not foolproof because it depends on the covariate (Rohrer, 2018). If the covariate is truly a confounder (C causes X, and C causes Y), then indeed the bivariate association between X and Y is spurious. However, if the covariate is a mediator (X causes C, and C causes Y), then controlling for the covariate removes the true causal association of interest and biases the findings. Moreover, if C is a collider (X causes C, and Y causes C) then controlling for the covariate can create a spurious association between X and Y. Sensitivity analyses can mitigate the issue of contradiction to some extent but only if these issues are given adequate attention.

Solutions

Having defined the overuse issue and delineated three specific problems that arise or are exacerbated by overuse, in our last section we discuss solutions. We link each of our proposed solutions to one or more of the three problems we had discussed above. In essence, what are practical steps researchers can take to ensure that even if available datasets see high use, analyses are carried out in a way that reduce false positives and other kinds of misleading findings? We divide our proposed solutions into two broad types: those that are self-imposed and those that are externally imposed. At the outset we recommend that self-imposed solutions are preferable as they minimize the possibility of regulatory overreach and are more likely to be met with acceptance from scholars in various fields.

Solutions Self-Imposed by Researchers: The Honor System

In general, we believe that restraints researchers impose on themselves to guard against the dangers of data overuse are best. These include greater data sharing, wider adoption of pre-registration, use of holdout samples, split-sample cross-validation, adjustment of alpha levels, entry of papers into dataset-based registries, and coordinated analysis. Some of these are described in Weston et al. (2019), but here we apply them specifically to the issue of data overuse.

Greater data sharing.—We anticipate that some may use the arguments made here as a justification to not share data or make data public. It is not hard to imagine someone indicating that they do not wish to add to the problem of data overuse, and so their data will remain inaccessible. Our answer to this is simple. The more data that is shared, the less of a problem overuse will be because the issue of scarcity will be reduced. Useful datasets that are not public or that are highly restricted add to the problem of overuse by forcing researchers toward the datasets that are public or more easily accessible. If more data were available there would be less dependence on a limited number of valuable studies that by happenstance are in the public domain.

This solution, greater sharing of data, directly addresses two of the three problems we discussed above: dependency and overfitting. The former would be ameliorated because by increasing available data, there would be a reduction in the number of findings in the literature that are reliant on a small set of datasets. Of course, some datasets are so valuable for the reason we mention above (large or diverse samples, many different types of variables, etc.) that some would still be more heavily used than others, keeping the dependency issue alive to some extent. Yet more data would reduce the problem. The overfitting and sampling error issue would also be mitigated because by placing more data in the public domain the peculiarities and quirks inherent in any one sample would be less likely to have an outsize impact.

Pre-registration.—The process of documenting research questions, hypotheses, and planned analyses prior to data analysis is not new, although it has received renewed focus in the past decade (Nosek et al., 2018), and new tools for pre-registering research in the social sciences have increased the popularity of this technique. While pre-registration has traditionally been conceptualized as a pre-data collection process, there is no reason that researchers using already-collected data should not preregister their analysis (Weston et al., 2019; see also the pre-registration template for existing data on the Open Science Foundation website). In the case of pre-existing data, researchers should take care to include information about prior knowledge of the dataset to be used. This should include both prior analyses using these data, as well as any known analyses by other researchers in the published literature. This brings us full circle with a point we made earlier when discussing the contradiction problem, namely that of reference limits. As part of our pre-registration solution, we also call for the lifting of caps on number of references, which may no longer be necessary given the increased use of digital articles. This permits extensive supplemental materials such as additional tables, full sets of materials and code, and data. References and bibliographies should be placed in the same category.

Among the aforementioned issues associated with overuse, the contradiction problem is the one directly impacted by pre-registration. Templates that exist for pre-registering investigations with existing data include sections where authors enumerate and describe prior published work on the relevant research questions using a given dataset. Preparing such a section allows authors to discover papers that otherwise would not have been found *before* analyzing the data. If the ultimate findings are at odds with a previous study, this would not come as a surprise and would be documented in the pre-registration. Of course, this solution does not preclude overusing certain datasets. Simply listing out prior findings would not necessarily curtail the heavy use of certain wide-used studies.

Use of holdout samples and cross-validation.—Most public datasets are large enough to allow splitting of the sample into random halves (or other splits such as 60%-40%) to determine if a given effect remains roughly the same across the two portions. Known as split-sample cross-validation, or rotation estimation (Stone, 1974), the approach involves separating the sample into a portion, called the training set, to develop a model. The remaining segment, known as the testing set, is used to validate. Many scholars have recently advocated use of cross-validation techniques to strengthen the robustness of our findings (e.g., Yarkoni et al., 2017).

A variation on this technique uses many random subsamples to generate a distribution of effect sizes (sometimes known as “repeated random subsampling cross-validation”). In this technique many subsamples, sometimes 10,000 or more, are used to estimate a mean effect of the many subsamples along with standards error. Other variations, “leave-p-out cross validation” and “leave-one-out cross validation,” are also in common use. Both involve leaving one or several subsamples out and then cross-validation on those.

Holdout samples with cross-validation specifically addresses the sampling error and overfitting problem. In fact, they can be a powerful hedge against this problem and thereby highlight the peculiarities that exist in any sample (Stone, 1974). These techniques can be powerful. However, in longitudinal applications where some participants are missing measurements at various waves, these cross-validation methods may require modification. In essence, each wave in a longitudinal study is a kind of subsample in and of itself due to study dropout, mortality, and other forms of attrition. Applying cross-validation techniques may require additional and careful thought when using them with longitudinal data. We also recognize that overuse of particular datasets may still occur even if holdout sampling becomes a regular practice.

Dataset-based registries.—As noted earlier, multiple publications on the same research questions using the same dataset is a common problem. Registries that are specific to a given dataset may be used to alleviate this issue. Such registries can collate and organize publications on a given research question that uses a given datasets, allowing researchers asking the same question to know exactly what has been done before. Ideally, this would prompt such researchers to first reproduce what earlier papers have found before adding new covariates, testing new interactions, breaking our subsamples for subgroup-specific analyses, or conducting analyses with a different statistical model. The existence and use of such registries would hopefully make researchers spell out, ideally in a pre-registration, exactly how their new analyses vary from prior ones on the same dataset.

Registries would also mitigate the problem of contradiction. A well-curated registry would make it easy for investigators to identify contradictory findings on the same research question using the same data. To be sure, such registries already exist for many public datasets, but often go unused. We recommend a new norm in individual study data sharing agreements (e.g., data use agreements; DUAs) that require researchers to have reviewed the published literature using that dataset and pledge to incorporate all prior findings into their introductions and analysis plans. This would be an honor system norm but is verifiable by other scholars and puts more accountability into the process.

Coordinated analysis.—Last, we endorse the wider use of coordinated analysis (Hofer et al., 2009; Mroczek, 2014). Coordinated analysis is form of integrative data analysis that involves the identification of multiple datasets having requisite data for answering a given question, estimating identical models across each dataset independently, and then synthesizing these results (e.g., Graham et al., 2017; see also other articles in this special issue). By design, coordinated analysis never relies on a single dataset. It therefore possesses built-in protection against two of the issues we described earlier. First and foremost, it addresses the sample peculiarity and overfitting problem. By not depending on a single dataset, coordinated analysis lessens the effect of any peculiarities inherent in a given sample and addresses the issue of overfitting. Using many samples in a coordinated analysis, for example 16 or more (e.g., Graham et al., 2020), can help to drown out the quirks that exist in each dataset, while also guarding against overfitting. This also has the effect of promoting generalizability. When many samples are used, especially if they are from diverse populations, then we can better distinguish between sampling error and true differences between populations. Coordinated analysis also addresses the contradiction problem because separate results are reported for all datasets. If some studies contradict one another, the contradictory results are out in the open and transparent. One important step of coordinated analysis is to identify datasets that have already published results on the association being addressed in the project. Including these datasets further provides information about the sensitivity and robustness of that effect to different analytic decisions, so long as the investigator discusses the results of the original publication. Additionally, while it’s true that the odds of false positives or contradictory findings increase with the overuse of a single data set, the nature of sampling error is such that spurious associations will disappear with aggregation across multiple samples. In a typical coordinated analysis (e.g., Graham et al., 2020), a wide distribution of effect sizes are observed across studies, including null effects, near-zero, and strong effects. Placing all of these estimates in context with one another provides a more reliable picture of what the “true” effect likely is, over and above any conclusion drawn from a single study. In this way, coordinated analysis helps reduce the consequences of data overuse.

Further, when coupled with our first solution, greater data sharing, coordinated analyses can become larger and more comprehensive. They can potentially make use of most of the available extant data needed for a given research question (e.g., consider an analysis of 83 extant datasets from around the world on alcohol and cardiovascular mortality; Wood et al., 2018). This promotes cumulative science while also reducing proliferation of studies on a given association using the same dataset, in turn addressing to some extent the dependency problem.

Solutions Imposed by External Bodies: Regulatory Systems

In other areas of resource management, especially of natural resources, when overuse has become a problem it has sometimes been necessary to impose external regulation. For example, governance bodies such as the Environmental Protection Agency enforce conservation and environmental laws. Natural resources and public resources (roadways, bridges) are subject to regulatory bodies and are also sometimes protected from overuse through fees or tolls. In scientific and medical research, Institutional Review Boards and Institutional Animal Care and Use Committees are examples of regulatory bodies. If the problems discussed here are not resolved by researcher self-imposed behavior change, it may at some point be necessary to invoke external regulatory bodies. We do not hope for this. Yet if needed, we have suggestions for how such regulations may be instituted.

Funding agencies can potentially set up in-house data overuse governance bodies that provide regulation. These agencies fund many of the studies that are made public, so it is natural for them to also engage in ways to ensure overuse of these resources does not inadvertently mislead scientists. They have a stake in ensuring the data they pay for is used in the best manner possible, even if that means occasionally restricting use for the greater good.

Many large ongoing studies have scientific advisory panels or data safety and monitoring boards. These bodies may also serve as regulators or gatekeepers of data. Approval of data use may be contingent upon some of the factors we described above, namely that potential users have conducted an exhaustive search of prior studies on similar research questions using that datasets and have made an argument as to how this new analysis will vary from previous ones. Preregistration should also be a pre-requisite for approval of data use. Of course, many public datasets already require data use agreements (DUAs). Thus, it would not be difficult to simply add to these existing DUA templates and insist on preregistration, a thorough prior literature search, and an explanation as to how the new analyses will be complementary or different from prior investigations as conditions of use.

If all of this seems heavy-handed and too restrictive of cherished notions of intellectual and scholarly freedom, we agree. Ideally, the situation will not come to require this. We perceive external regulation as a last resort, and we would prefer to see researchers adopt at least some of the many solutions offered in the self-regulation section. Self-imposed regulation is better than the inevitable bureaucratic and administrative hurdles that would accompany any sort of external solution. One possibility that may help avoid externally-imposed regulation is some combination of self-imposed behavior that are coupled with “in-between” levels of authority, such as journal editors and article reviewers. Self-imposed solutions combined with changed journal expectations and editorial policies may help avoid burdensome external regulation.

We are hopeful. The response to the replication and credibility crisis in multiple scientific fields has led to new norms and mores among scientists (particularly younger scholars and early-career researchers) without any formal regulatory bodies being created. Self-imposition seems to have worked for the replication and credibility movements and our fervent hope is that it will work for the problem of data overuse.

Conclusions

The value of many public datasets is being diminished by overuse. Like other public resources that are overused, there are different ways of handling this dilemma. Some involve self-imposed norms by individual researchers or research teams and others rely on external regulatory bodies. We prefer the former. In doing so, we may yet avoid the tragedy of the commons in our own field.

Acknowledgments:

We thank Julia Rohrer for useful comment on an earlier version of this paper.

Funding:

This research was supported by grants from the National Institute on Aging: R01-AG018436 (Personality & Well-being Trajectories in Adulthood) and R01-AG067622 (Personality and Prediction of Dementia Risk & Progression)

References

Babyak MA (2004). What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models. Psychosomatic Medicine, 66(3), 411–421. 10.1097/01.psy.0000127692.23278.a9 [PubMed: 15184705]

Caspi A Sugden K Moffitt TE, Taylor A, Craig IW, Harrington HL, McClay J, Mill J, Judy Martin J, Braithwaite A, Poulton R (2003). Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene. Science, 301, 386–389. DOI: 10.1126/science.1083968 [PubMed: 12869766]

Chatterjee S, & Hadi AS (2009). Sensitivity analysis in linear regression (Vol. 327). John Wiley & Sons.

Condon DM, Graham EK, & Mroczek DK (2018). On replication research. PsyArXiv. 10.31234/osf.io/2fn5x

Ferrie JP, Rolf K & Troesken W (2012).Cognitive disparities, lead plumbing, and water chemistry: Prior exposure to water-borne lead and intelligence test scores among World War Two U.S. Army enlistees. Economics & Human Biology, 10, 98–111. doi: 10.1016/j.ehb.2011.09.003. [PubMed: 22014834]

Graham EK, & Lachman ME (2012). Personality stability is associated with better cognitive performance in adulthood: Are the stable more able?. Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 67(5), 545–554. 10.1093/geronb/gbr149

Graham EK, Rutsohn JP, Turiano NA, Bendayan R, Batterham P, Gerstorf D, Katz M, Reynolds C, Schoenhofen E, Yoneda T, Bastarache E, Elleman LG, Zelinski EM, Johansson B, Kuh D, Barnes LL, Bennett D, Deeg D, Lipton R, Pedersen N, Piccinin A, Spiro A, Muniz-Terrera G, Willis S, Schaie KW, Roan C, Herd P, Hofer SM, & Mroczek DK (2017). Personality predicts mortality risk: An integrative analysis of 15 international longitudinal studies. Journal of Research in Personality, 70, 174–186. [PubMed: 29230075]

Graham EK, Weston SJ, Gerstorf D, Yoneda TB, Booth T, Beam CR, Petkus AJ, Drewelies J, Hall AN, Bastarache ED, Estabrook R, Katz MJ, Turiano NA, Lindenberger U, Smith J, Wagner GG, Pedersen NL, Allemand M, Spiro A, Deeg DJH, Johansson B, Piccinin AM, Lipton RB, Schaie KW, Willis S, Reynolds CR, Deary IJ, Hofer SM, & Mroczek DK (2020). Trajectories of Big Five Personality Traits: A coordinated Analysis of 16 Longitudinal Samples. European Journal of Personality, 34, 301–321. DOI :10.1002/per.2259 [PubMed: 33564207]

Henrich J, Heine SJ, & Norenzayan A (2010). Most people are not WEIRD. Nature, 466(7302), 29–29. [PubMed: 20595995]

Hofer SM, & Piccinin AM (2009). Integrative data analysis through coordination of measurement and analysis protocol across independent longitudinal studies. Psychological Methods, 14(2), 150–164. [PubMed: 19485626]

Low L-F, Harrison F, & Lackersteen SM (2013). Does personality affect risk for dementia? A systematic review and meta-analysis. American Journal of Geriatric Psychiatry, 21, 713–728. 10.1016/j.jagp.2012.08.004

Mroczek DK, Pitzer LM, Miller LM, Turiano NA, & Fingerman KF (2011). The use of secondary data in adult development and aging research. In Trzesniewski KH, Donnellan MB & Lucas RE (Eds.), Secondary data analysis: An introduction for psychologists (pp., 121–132). Washington, DC: American Psychological Association.

Mroczek DK (2014). Personality plasticity, healthy aging, and interventions. Developmental Psychology, 50, 1470–1474. [PubMed: 24773109]

Munafò MR, Durrant C, Lewis G, & Flint J (2009). Gene× environment interactions at the serotonin transporter locus. Biological psychiatry, 65(3), 211–219. [PubMed: 18691701]

Nosek BA, Ebersole CR, DeHaven AC, & Mellor DT (2018). The preregistration revolution. Proceedings of the National Academy of Sciences, 115(11), 2600–2606.

Risch N, Herrell R, Lehner T, Liang KY, Eaves L, Hoh J, … & Merikangas KR (2009). Interaction between the serotonin transporter gene (5-HTTLPR), stressful life events, and risk of depression: a meta-analysis. JAMA, 301(23), 2462–2471. [PubMed: 19531786]

Rohrer JM (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. Advances in Methods and Practices in Psychological Science, 1(1), 27–42

Saltelli A, Tarantola S, Campolongo F, & Ratto M (2004). Sensitivity analysis in practice: a guide to assessing scientific models (Vol. 1). New York: Wiley.

Simons DJ, Shoda Y, & Lindsay DS (2017). Constraints on generality (COG): A proposed addition to all empirical papers. Perspectives on Psychological Science, 12(6), 1123–1128. [PubMed: 28853993]

Simmons JP, Nelson LD, & Simonsohn U (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological science, 22(11), 1359–1366. [PubMed: 22006061]

Stone M (1974). Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society, 36, 111–133. 10.1111/j.2517-6161.1974.tb00994.x

Sutin AR, Stephan Y, Damian RI, Luchetti M, Strickhouser JE, & Terracciano A (2019). Five-factor model personality traits and verbal fluency in 10 cohorts. Psychology and aging, 34(3), 362. 10.1037/pag0000351 [PubMed: 31070400]

Thompson WH, Wright J, Bissett PG, & Poldrack RA (2020). Dataset decay and the problem of sequential analyses on open datasets. eLife, 9, e53498. doi: 10.7554/eLife.53498 [PubMed: 32425159]

Weston SJ, Ritchie SJ Rohrer JM & Przybylski AK (2019). Recommendations for Increasing the Transparency of Analysis of Preexisting Data Sets. Advances in Methods and Practices in Psychological Science, 2(3), 214–227. DOI: 10.1177/2515245919848684 [PubMed: 32190814]

Wicherts JM, Veldkamp CL, Augusteijn HE, Bakker M, Van Aert R, & Van Assen MA (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. Frontiers in psychology, 7, 1832. [PubMed: 27933012]

Wood AM, Kaptoge S, Butterworth AS et al. (2018). Risk thresholds for alcohol consumption: combined analysis of individual-participant data for 599,912 current drinkers in 83 prospective studies. Lancet, 391, 1513–1523. [PubMed: 29676281]

Yarkoni T (in press). The generalizability crisis. Brain & Behavioral Sciences.

Yarkoni T & Westfall J (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. Perspectives on Psychological Science, 12, 1100–1122. https://journals.sagepub.com/doi/full/10.1177/1745691617693393 [PubMed: 28841086]