Research and Reviews

1. Marijuana: Much More Than You Wanted To Know
2. Wheat: Much More Than You Wanted To Know
3. SSRIs: Much More Than You Wanted To Know
4. Alcoholics Anonymous: Much More Than You Wanted To Know
5. Prescriptions, Paradoxes, and Perversities
6. Guns And States
7. Teachers: Much More Than You Wanted To Know
8. Antidepressant Pharmacogenomics: Much More Than You Wanted To Know
9. A Story With Zombies
10. Asches to Asches

Marijuana: Much More Than You Wanted To Know

This month I work on my hospital's Substance Abuse Team, which means we treat people who have been hospitalized for alcohol or drug-related problems and then gingerly suggest that maybe they should use drugs a little less.

The two doctors leading the team are both very experienced and have kind of seen it all, so it's interesting to get a perspective on drug issues from people on the front line. In particular, one of my attendings is an Obama-loving long-haired hippie who nevertheless vehemently opposes medical marijuana or any relaxation on marijuana's status at all. He says that "just because I'm a Democrat doesn't mean I have to support stupid policies I know are wrong" and he's able to back up his opinion with an impressive variety of studies.

To be honest, I had kind of forgotten that the Universe was allowed to contain negative consequences for legalizing drugs. What with all the mental energy it took protesting the the Drug War and getting outraged at police brutality and celebrating Colorado's recently permitting recreational cannabis use and so on, it had completely slipped my mind that the legalization of marijuana might have negative consequences and that I couldn't reject it out of hand until I had done some research.

So I've been doing the research. Not to try to convince my attending of anything – as the old saying goes, do not meddle in the affairs of attendings, because you are crunchy and taste good with ketchup – but just to figure out where exactly things stand.

I. Would Relaxation Of Penalties On Marijuana Increase Marijuana Use?

Starting in the 1970s, several states decriminalized possession of marijuana – that is, possession could not be penalized by jail time. It could still be penalized by fines and other smaller penalties, and manufacture and sale could still be punished by jail time.

Starting in the 1990s, several states legalized medical marijuana. People with medical marijuana cards, which in many cases were laughably easy to get with or without good evidence of disease, were allowed to grow and use marijuana, despite concerns that some of this would end up on the illegal market.

Starting last week, Colorado legalized recreational use of marijuana, as well as cultivation and sale (subject to heavy regulations). Washington will follow later this year, and other states will be placing measures on their ballots to do the same.

One should be able to evaluate to what degree marijuana use rose after these policy changes, and indeed, many people have tried – with greater or lesser levels of statistical sophistication.

The *worst* arguments in favor of this proposition are those like this CADCA paper, which note that states with more liberal marijuana laws have higher rates of marijuana use among teenagers than states that do not. The proper counterspell to such nonsense is *Reverse Causal Arrows* – could it not be that states with more marijuana users are more likely to pass proposals liberalizing marijuana laws? Yes it could. Even more likely, some third variable – let's call it "hippie attitudes" – could be

behind both high rates of marijuana use and support for liberal marijuana regimes. The states involved are places like Colorado, California, Washington, and Oregon. I think that speaks for itself. In case it doesn't, someone went through the statistics and found that these states had the highest rates of marijuana use among teens since *well* before they relaxed drug-related punishments. Argument successfully debunked.

A slightly more sophisticated version – used by the DEA here – takes the teenage marijuana use in a state one year before legalization of medical marijuana and compares it to the teenage marijuana use in a state one (or several years) after such legalization. They often find that it has increased, and blame the increase on the new laws. For example, 28% of Californians used marijuana before it was decriminalized in the 70s, compared to 35% a few years after. This falls victim to a different confounder – marijuana use has undergone some very large swings nationwide, so the rate of increase in medical marijuana states may be the same as the rate anywhere else. Indeed, this is what was going on in California – its marijuana use actually rose slightly *less* than the national average.

What we want is a study that compares the average marijuana use in a set of states before liberalization to the average marijuana use in the country as a whole, and then does the same after liberalization to see if the ratio has increased. There are several studies that purport to try this, of which by far the best is Johnston, O'Malley & Bachman 1981, which monitored the effect of the decriminalization campaigns of the 70s. They survey thousand of high school seniors on marijuana use in seven states that decriminalize marijuana both before and for five years after the decriminalization, and find absolutely no sign of increased marijuana use (in fact, there is a negative trend). Several other studies (eg Thies & Register 1993) confirm this finding.

There is only a hint of some different results. Saffer and Chaloukpa 1999 and Chaloupka, Grossman & Tauras 1999 try to use complicated econometric simulations to estimate the way marijuana demand will respond to different variables. They simulate (as opposed to detecting in real evidence) that marijuana decriminalization should raise past-year use by about 5 – 8%, but have no effect on more frequent use (ie a few more people try it but do not become regular users). More impressively, Model 1993 (a source of some exasperation for me earlier) finds that after decriminalization, marijuana-related emergency room visits went up (trying to interpret their tables, I think they went up by a whopping 90%, but I'm not sure of this). This is sufficiently different from every other study that I don't give it much weight, although we'll return to it later.

Overall I think the evidence is pretty strong that decriminalization probably led to no increase in marijuana use among teens, and may at most have led to a small single-digit increase.

Proponents of stricter marijuana penalties say the experiment isn't fair. In practice, decriminalization does not affect the average user very much – even in states without decriminalization, marijuana possession very rarely leads to jail time. The only hard number I have is from Australia, where in "non-decriminalized" Australian states only 0.3% of marijuana arrests lead to jail time, but a quick back-of-the-envelope calculation suggests US numbers are very similar. And even in supposedly decriminalized states, it's not hard for a cop who wants to get a pot user in jail to find a way (possession of even small amounts can be "possession with intent to sell" if someone doesn't like you). So the overall real difference between decriminalized and not decriminalized is small and it's not surprising the results are small as well. I mostly

agree with them; decriminalization is fine as far as it goes, but it's a bigger psychological step than an actual one.

The next major milestone in cannabis history was the legalization of medical marijuana. Anderson, Hansen & Rees (2012) did the same kind of study we have seen above, and despite trying multiple different measures of youth marijuana use found pretty much no evidence that medical marijuana legalization caused it to increase. Other studies find pretty much the same.

This could potentially suffer from the same problems as decriminalization studies – the laws don't always change the facts on the ground. Indeed, for about ten years after medical marijuana legalization, the federal government kept on prosecuting marijuana users even when their use accorded with state laws, and many states had so few dispensaries that in reality not a whole lot of medical marijuana was being given out. I haven't found any great studies that purport to overcome these problems.

When we examined decriminalization, we found that the studies based on surveys of teens looked pretty good, but that the one study that examined outcomes – marijuana-related ER visits – was a lot less encouraging. We find the same pattern here, and the rain on our parade is Chu 2013, who finds that medical marijuana laws increased marijuana-related arrests by 15-20% and marijuana-related drug rehab admissions by 10-15%.

So what's going on here? I have two theories. First, maybe medical marijuana use (and decriminalization) increase use among adults only. This could be because the system is working – giving adults access to medical marijuana while keeping it out of the hands of children – or because kids are dumb and don't understand consequences but adults are more responsive to incentives and punishments. Second, we know that medical marijuana has twice as much THC as street marijuana. Maybe everyone keeps using the same amount of marijuana, but when medical marijuana inevitably gets diverted to the street, addicts can't handle it and end up behaving much worse than they expected.

Or the studies are wrong. Studies being wrong is always a pretty good bet.

I can't close this section without mentioning the Colorado expulsion controversy. Nearly everyone who teaches in Colorado says there has been an explosion of marijuana-related problems since medical marijuana was legalized. Meanwhile, the actual surveys of Colorado high school students say that marijuana use, if anything, is going down. A Colorado drug warrior has some strong objections to the survey results, but they center around not really being able to prove that there is a real downward trend (which is an entirely correct complaint) without denying that in fact they show no evidence at all of going *up*.

The consensus on medical marijuana seems to be that it does not increase teen marijuana use either, although there is some murky and suggestive evidence that it might increase illicit or dangerous marijuana use among adults.

There is less information on the effects of full legalization of marijuana, which has never been tried before in the United States. To make even wild guesses we will have to look at a few foreign countries plus some econometric simulations.

No one will be surprised to hear that the first foreign country involved is the Netherlands, which was famously permissive of cannabis up until a crackdown a few years ago. Despite popular belief they never fully legalized the drug and they were

still pretty harsh on production and manufacture; distribution, on the other hand, could occur semi-openly in coffee shops. This is another case where we have to be careful to distinguish legal regimes from actual effects, but during the period when there were actually a lot of pot-serving coffee shops, the Netherlands did experience an otherwise-inexplicable 35% rise in marijuana consumption relative to the rest of Europe. This is true even among teenagers, and covers both heavy use as well as occasional experimentation. Some scientists studying the Netherlands' example expect Colorado to see a similar rise; others think it will be even larger because the legalization is complete rather than partial.

The second foreign country involved is Portugal, which was maybe more of a decriminalization than a legalization case but which is forever linked with the idea of lax drug regimes in the minds of most Americans. They decriminalized all drugs (including heroin and cocaine) in 2001, choosing to replace punishment with increased treatment opportunities, and as we all have been told, no one in Portugal ever used drugs ever again, or even remembers that drugs exist. Except it turns out it's more complicated; for example, the percent of Portuguese who admit to lifetime use of drugs has doubled since the law took effect. Two very patient scientists have sifted through all the conflicting claims and found that in reality, the number of people who briefly experiment with drugs has gone way up, but the number of addicts hasn't, nor has the number of bad outcomes like overdose-related deaths. There are many more people receiving drug treatment, but that might just be because Portugal upped its drug treatment game in a separate law at the same time they decriminalized drugs. Overall they seem to have been a modest success – neither really raising nor decreasing the number of addicts – but they seem more related to decriminalization (which we've already determined doesn't have much effect) than to legalization per se.

Returning to America, what if you just *ask* people whether they would use more marijuana if it's legal? Coloradans were asked if they plan to smoke marijuana once it becomes legal; comparing survey results to current usage numbers suggests 40% more users above the age of 18; it is unclear what the effect will be on younger teens and children.

Finally, we let the economists have their say. They crunch all the data and predict an increase of 50 – 100% based solely on the likely price drop (even with taxes factored in). And if there's one group we can trust to make infallible predictions about the future, it's economists.

Overall I find the Dutch evidence most convincing, and predict a 25 – 50% increase in adult marijuana use with legalization. I would expect a lower increase – 15 – 30% – among youth, but the data are also perfectly consistent with no increase at all.

Conclusion for this section: that decriminalization and legalization of medical marijuana do not increase youth marijuana use rates, although there is some shaky and indirect evidence they do increase adult use and bad behavior. There is no good data yet on full legalization, but there's good reason to think it would substantially increase adult use and it might also increase youth use somewhat.

II. Is Marijuana Bad For You?

About 9% of marijuana users eventually become addicted to the drug, exposing them to various potential side effects.

Marijuana smoke contains a lot of the same chemicals in tobacco smoke and so it would not be at all surprising if it had some of the same ill effects, like cardiovascular disease and lung cancer. But when people look for these effects, they can't find any increase in mortality among marijuana smokers. I predict that larger studies will one day pick something up, but for now let's take this at face value.

Much more concerning are the attempts to link marijuana to cognitive and psychiatric side effects. Meier et al (2012) analyzed a study of a thousand people in New Zealand and found that heavy marijuana use was linked to an IQ decline of 8 points. Rogeberg 2012 developed an alternative explanation – poor people saw their IQs drop in their 20s more than rich people because their IQs had been artificially inflated by schooling; what Meier et al had thought to be an effect of cannabis was really an effect of poor people having an apparent IQ drop and using cannabis more often. Meier et al pointed out that actually, poor people didn't use cannabis any more often than anyone else and effects remained when controlled for class. Other studies, like Fried et al (2002) find the same effect, and there is a plausible biological mechanism (cannabinoids something something neurotransmitters something brain maturation). As far as I can tell the finding still seems legit, and marijuana use does decrease IQ. It is still unclear whether this only applies in teenagers (who are undergoing a "sensitive period of brain development") or full stop.

More serious still is the link with psychosis. A number of studies have found that marijuana use is heavily correlated with development of schizophrenia and related psychotic disorders later in life. Some of them find relative risks as high as 2 – heavy marijuana use doubles your chance of getting schizophrenia, which is already a moderately high 1%. But of course correlation is not causation, and many people have come up with alternative theories. For example, maybe people who are already kind of psychotic use marijuana to self-medicate, or just make poor life choices like starting drugs. Maybe people of low socioeconomic status who come from broken homes are more likely to both use marijuana and get schizophrenia. Maybe some gene both makes marijuana really pleasant and increases schizophrenia risk.

I know of three good studies attempting to tease out causation. Arseneault et al (2004) checks to see which came first – the marijuana use or the psychotic symptoms – and finds it was the marijuana use, thus supporting an increase in risk from the drug. Griffith-Lendering et al (2012) try the same, and find *bidirectional* causation – previous marijuana use seems to predict future psychosis, but previous psychosis seems to predict future marijuana use. A very new study from last month boxes clever and checks whether your marijuana use can predict schizophrenia in *your relatives*, and find that it does – presumably suggesting that genetic tendencies towards schizophrenia cause marijuana use and not vice versa (although Ozy points out to meet that the relatives of marijuana users are more likely to use marijuana themselves; the plot thickens). When a meta-analysis tries to control for all of these factors, they get a relative risk of 1.4 (they call it an odds ratio, but from their discussion section I think they mean relative risk).

Is this true, or just the confounders they failed to pick up? One argument for the latter is that marijuana use has increased very much over the past 50 years. If marijuana use caused schizophrenia, we would expect to see much more schizophrenia, but in fact as far as anyone can tell (which is not very far) schizophrenia incidence is decreasing. The decrease might be due (maybe! if it even exists at all!) to obstetric advances which prevent fetal brain damage which could later lead to the disease. The effect of this variable is insufficiently known to pretend we can tease out some

supposed contrary effect of increased marijuana use. Also, some people say that schizophrenia is increasing in young people, so who knows?

The *exact* nature of the marijuana-psychosis link is still very controversial. Some people say that marijuana causes psychosis. Other people say it "activates latent psychosis", a term without a very good meaning but which might mean that it pushes people on the borderline of psychosis – eg those with a strong family history but who might otherwise have escaped – over the edge. Still others say all it does is get people who would have developed psychosis eventually to develop it a few years earlier. You can read a comparison of all the different hypotheses here.

I've saved the most annoying for last: is marijuana a "gateway drug"? Would legalizing it make it more or less of a "gateway drug"? This claim seems tailor-made to torture statisticians. We know that marijuana users are *definitely* more likely to use other drugs later – for example, marijuana users are 85x more likely than non-marijuana users to use cocaine. but that could be either because marijuana affects them in some way (implying that legalizing marijuana would increase other drug use), because they have factors like genetics or stressful life situation that makes them more likely to use all drugs (implying that legalizing marijuana would not affect other drug use), or because using illegal marijuana without ill effect connects them to the illegal drug market and convinces them illegal drugs are okay (implying that legalizing marijuana would decrease other drug use). RAND comes very close to investigating this properly by saying that when the Dutch pseudo-legalized marijuana, use of harder drugs stayed stable or went down, but all their study actually shows is that the ratio of marijuana users : hard drug users went down. This is to be expected when you make marijuana much easier to get, but it's still consistent with the absolute number of hard drug users going way up. The best that can be said is that there is no direct causal evidence for the gateway theory and some good alternative explanations for the effect. Let us accept their word for it and never speak of this matter again.

Conclusion for this section: Marijuana does not have a detectable effect on mortality and there is surprisingly scarce evidence of tobacco-like side effects. It probably does decrease IQ if used early and often, possibly by as many as 8 IQ points. It may increase risk of psychosis by as much as 40%, but it's not clear who is at risk or whether the risk is even real. The gateway drug hypothesis is too complicated to evaluate effectively but there is no clear casual evidence in its support.

III. What Are The Costs Of The Drug War?

There are not really that many people in jail for using marijuana.

I learned this from Who's Really In Prison For Marijuana?, a publication of the National Office Of Drug Control Policy, which was clearly written by someone with the same ability to take personal offense at bad statistics that inspires my posts about Facebook. The whole thing seethes with indignation and makes me want to hug the drug czar and tell him everything will be okay.

Only 1.6% of state prisoners are serving time for marijuana, only 0.7% are serving for marijuana possession, and only 0.3% are first time offenders. Some of those are "possession" in the sense of "possessing a warehouse full of marijuana bales", and others are people who committed much more dangerous crimes but were nailed for marijuana, in the same sense that Al Capone was nailed for tax evasion. The percent of normal law-abiding people who just had a gram or two of marijuana and were

thrown in jail is a rounding error, and the stories of such you read in the news are extremely dishonest (read the document for examples).

Federal numbers are even lower; in the entire federal prison system, they could only find 63 people imprisoned with marijuana possession as the sole crime, and those people were possessing a median of one hundred fifteen *pounds* of marijuana (enough to make over 100,000 joints).

In total, federal + state prison and counting all the kingpins, dealers, manufacturers, et cetera, there are probably about 16,000 people in prison solely for marijuana-related offenses, serving average actual sentence lengths of three year. But it's anybody's guess whether those people would be free today if marijuana were legal, or whether their drug cartels would just switch to something else.

Looking at the other side's statistics, I don't see much difference. NORML claims that there are 40,000 people in prison for marijuana use, but they admit that half of those people were arrested for using harder drugs and marijuana was a tack-on charge, so they seem to agree with the Feds about around 20,000 pure marijuana prisoners. SAM agrees that only 0.5% of the prison population is in there for marijuana possession alone. I see no reason to doubt any of these numbers.

A much more serious problem is marijuana-related arrests, of which there are 700,000 a year. 90% of them are for simple possession, and the vast majority do not end in prison terms; they do however result in criminal records, community service, a couple days of jail time until a judge is available to hear the case, heavy fines, high cost of legal representation, and moderate costs to the state for funding the whole thing. Fines can be up to $1500, and legal representation can cost up to $5000 (though I am suspicious of this paper and think it may be exaggerating for effect). These costs are often borne by poor people who will have to give up all their savings for years to pay them back.

Costs paid by the government, which cover everything from police officers to trials to prison time, are estimated at about $2 billion by multiple sources. This is only 3% of the total law enforcement budget, so legalizing marijuana wouldn't create some kind of sudden revolution in policing, but as the saying goes, a billion here, a billion there, and eventually it adds up to real money. And a Harvard economist claims that the total monetary benefits from legalization, including potential tax revenues, could reach $14 billion.

Some people worry that legalizing marijuana would cause an increase in car accidents by "stoned drivers", who, like drunk drivers, have impaired reflexes and poor judgment, and indeed there is a small but real problem of marijuana-induced car accidents. But Chaloukpa and Laixuthai (1994) crunch the numbers and find that decreased price/increased availability of marijuana is actually associated with *decreased* car accidents, probably because marijuana is substituting for alcohol in the "have impairing substances and then go driving" population. This finding – that marijuana and alcohol substitute for each other – has been spotted again and again. Anderson & Rees (2013) find that states that legalize medical marijuana see a 5% drop in beer sales. There are however a few dissenting opinions: Cameron & Williams (2001), in complex econometric simulations that may or may not resemble the real world in any respect, find that increasing the price of alcohol increases marijuana use, but increasing the price of marijuana does not affect alcohol use, and the same researcher finds that banning alcohol on a college campus also decreases marijuana use. Also, possibly marijuana use increases smoking? This whole area is confusing, but

I am most sympathetic to to the Andersen and Rees statistics which say that medical marijuana states are associated with 13% fewer traffic fatalities.

Overall conclusion for this section: full legalization of marijuana would free about 20,000 people from jail (although most of them would not be exactly fine upstanding citizens), prevent 700,000 arrests not resulting in jail time per year, save between 2 and 14 billion dollars, and possibly reduce traffic fatalities a few percent (or, for all we know, increase them).

IV. An Irresponsible Utilitarian Analysis

Decriminalization and legalization of medical marijuana seem, if we are to trust the statistics in (I) saying they do not increase use among youth, like almost unalloyed good things. Although there are some nagging hints of doubt, they are not especially quantifiable and therefore not amenable to analysis. Without a very strong predisposition to try as hard as possible to fit the evidence into a pessimistic picture, I don't think there's a great argument against either of these two propositions. Let's concentrate on legalization, which would mean something like "People can grow and sell as much marijuana as they want and it's totally legal for people over 21, with the same level of penalties as today for people under 21".

Section (I) concludes that legalization could lead to an increase in adult marijuana use up to 50%. There's not a lot of evidence on what it could do to teen marijuana use, but since it seems teen marijuana use is less responsive to legal changes, I made up a number and said 20%. Lest you think I am being unfair, note that this is well below the percent increase predicted by the survey that asked 18 year olds if they would start using marijuana if it were legal.

Right now about 1.5 million teenagers use marijuana "heavily". Most of the detrimental effects of marijuana seem concentrated in teens and people in their early twenties; I'm going to artificially round that up to 2 million to catch the early 20 year olds. If this 2 million number increased 20%, 400,000 extra teens would start heavily using marijuana.

Those 400,000 teens would lose 8 IQ points each. IQ increases your yearly earnings by about $500 per point, so these people would lose about $4,000 a year. Making very strong assumptions about salary being a measure of value to society, society would lose about $1.6 billion a year directly, plus various intangibles from potential artists and scientists losing the ability to create masterpieces and inventions, plus various *really* intangibles like a slightly dumber electorate.

We need to use a different number to calculate psychosis risk, since the studies were done on "people who had used marijuana at least once". The appropriate number turns out to be 8 million teenagers; of those, 1%, or 80,000, would naturally develop schizophrenia. If the 1.4 relative risk number is correct, marijuana use will increase that to 112,000, for a total increase of 32,000 people. Schizophrenia pretty much always presents in the 15 – 25 age window, so we'll say we get 3,200 extra cases per year.

There were 35000 road traffic accident fatalities in the US last year. If greater availability of marijuana decreases those fatalities by 13% (note that I am using the number from medical marijuana legalization and not for marijuana legalization per se, solely because it is a number I actually have), that will cause 4500 fewer road traffic deaths per year. There may be additional positive effects of alcohol substitution from,

for example, less liver disease. But there may also be additional negative effects from increasing use of tobacco, so let's just pretend those cancel out.

So here is my guess at the yearly results of marijuana legalization:

– 20,000 fewer prisoners (but they might switch to other criminal enterprises)
– 700,000 fewer arrests
– $2 billion less in law enforcement costs
– Some amount of positive gain (let's say $5 billion) in taxes
– 4500 fewer road traffic deaths (if you believe the preliminary alcohol substitution numbers)

– 400,000 people with lower IQ
– $2 billion in social costs from above dumber people
– 3,200 more cases of schizophrenia a year

We'll proceed to calculate the nonmonetary burden of each of these in QALYs, then add the monetary burden in dollars, then convert.

The searchable public database of utility weights for all diseases (God I love the 21st century) tells me that schizophrenia has a QALY weight of 0.73. It generally starts around 20 and lasts a lifetime, so each case of schizophrenia costs us 0.27 * 50 or 13.5 QALYs. Therefore, the total burden of the 3,200 added schizophrenia cases is 43 kiloQALYs.

There's no good way to calculate the QALY weight of having 4-8 fewer IQ points, and unfortunately this is going to end up being among the most important numbers in our results. If we say the lifetime cost of this problem is 3 QALYs, and divide the number by eight to represent eight years worth of teenagers in our sample population, we end up with 400,000/8 * 3 = 150 kiloQALYs.

My own survey tells me that being in prison has a QALY weight around 0.5. Marijuana sentences generally last an average of three years, which suggests that 1/3 of these marijuana prisoners are arrested every year, so the total burden of the ~6000ish marijuana imprisonments each year is 3 * ~6000 * 0.5 = 10 kiloQALYs.

Assume the average road traffic death occurs at age 30, costing 40 years of potential future life. The total cost of 4500 road traffic deaths is 40 * 4500 = 180 kiloQALYs.

The arrests are going to require even more fudging than normal. Average jail time for a marijuana arrest (when awaiting trial) is "one to five days" – let's round that off to two and then use our prison number to say that the jail from each arrest is 2/365 * 0.5 = three-thousandths of a QALY. I am going to arbitrarily round this up to one one-hundredth of a QALY to account for emotional trauma and the burden of fines, then even more arbitrarily round this up to a tenth of a QALY to account for possibility of getting a criminal record. This sets the burden of 700,000 arrests at 70 kiloQALYs.

Now our accounting is:

Costs from legalization compared to current system: 200 kQALYs and $2 billion
Benefits from legalization compared to current system: 260 kQALYs and $7 billion

Although it's not going to be necessary, we can interconvert QALYs and dollars at the going health-care rate of about $100,000/QALY ($100 million/kQALY):

Costs from legalization compared to current system: 220 kQALYs
Benefits from legalization compared to current system: 330 kQALYs

And get:

Net benefits from legalization: +110 kQALYs

Except that this is extremely speculative and irresponsible. By far the largest component of the benefits of legalization turned out to be the effect on road traffic accidents, which is based on only two studies and which may on further research turn out to be a cost. And by far the largest component of the costs of legalization turned out to be the effect on IQ, and we had to totally-wild-guess the QALY cost of an IQ point loss. The wiggle room in my ignorance and assumptions is more than large enough to cover the small gap between the two policies in the results.

So my actual conclusion is:

There is not a sufficiently obvious order-of-magnitude difference between the costs and benefits of marijuana legalization for a evidence-based utilitarian analysis of costs and benefits to inform the debate. You may return to your regularly scheduled wild speculation and shrill accusations.

But I wouldn't say this exercise is useless. For example, it suggests that whether marijuana legalization is positive or negative on net depends almost entirely on small changes in the road traffic accident rate. This is something I've never heard anyone else mention, but which in retrospect should be obvious; the few debatable health effects and the couple of people given short jail sentences absolutely can't compare to the potential for thousands more (or fewer) traffic accidents which leave people permanently dead.

So my actual actual conclusion is:

We should probably stop caring about health effects of marijuana and about imprisonment for marijuana-related offenses, and concentrate all of our research and political energy on how marijuana affects driving.

This cements my previous intuitions on irresponsible use of statistics – it's unlikely to unilaterally solve the problem, but it can be very good at pointing out where you're being irrational and suggesting new ways of looking at a question.

EDIT: People in the comments have pointed out several important factors left out, including:
– Some people enjoy smoking marijuana
– The opening of a permanent criminal record may mean arrests are worse than I estimate. I can't find good statistics on how often this happens, but do note that decriminalization prevents a record from being opened.
– Loss of 8 IQ points may have wider social effects than I estimate, since IQ affects for example crime rate.
– Legalizing marijuana might remove a source of funding for organized crime

Wheat: Much More Than You Wanted To Know

After hearing conflicting advice from diet books and the medical community, I decided to look into wheat.

There are two sets of arguments against including wheat in the diet. First, wheat is a carbohydrate, and some people support low carbohydrate diets. Second, something might be especially dangerous about wheat itself.

It was much easier to figure out the state of the evidence on low-carbohydrate diets. They seem to be at least as good and maybe a little better for weight loss than traditional diets, but this might just be because there are lots of carbohydrates that taste very good and when forced to avoid them, people eat less stuff. They may or may not positively affect metabolic parameters and quality of life. (1, 2, 3, 4). They don't seem to cause either major health benefits or major health risks in the medium term, which is the longest term for which there is good data available – for example, they have no effect on cancer rates. Overall they seem solid but unspectacular. But there's a long way between "low carbohydrate diet" and "stop eating wheat".

So I was more interested in figuring out what was going on with wheat in particular.

Wheat contains chemicals [citation needed]. The ones that keep cropping up (no pun intended) in these kinds of discussions are phytates, lectins, gluten, gliadin, and agglutinin, the last three of which for your convenience have been given names that all sound alike.

Various claims have been made about these chemicals' effects on health. These have some prima facie plausibility. Plants don't want to be eaten [citation needed] and they sometimes fill their grains with toxins to discourage animals from eating them. Ricin, a lectin in the seeds of the castor oil plant so toxic it gets used in chemical warfare, is a pretty good example. Most toxins are less dramatic, and most animals have enzymes that break down the toxins in their preferred food sources effectively. But if humans are insufficiently good at this, maybe because they didn't evolve to eat wheat, some of these chemicals could be toxic to humans.

On the other hand, this same argument covers every pretty much every grain and vegetable and a lot of legumes – pretty much every plant-based food source except edible fruits. So we need a lot more evidence to start worrying about wheat.

I found the following claims about negative effects of wheat:

1. Some people without celiac disease are nevertheless sensitive to gluten.
2. Wheat increases intestinal permeability, causing a leaky gut and autoimmune disease.
3. Digestion of wheat produces opiates, which get you addicted to wheat.
4. Wheat something something something autism and schizophrenia.
5. Wheat has been genetically modified recently in ways that make it much worse for you.
6. The lectins in wheat interfere with leptin receptors, making people leptin resistant and therefore obese.

I'll try to look at each of those and then turn to the positive claims made about wheat to see if they're strong enough to counteract them.

Some People Without Celiac Disease Are Sensitive To Gluten – *Mostly true but of limited significance*

Celiac disease is one source of concern. Everybody on all sides of the wheat debate agree about the basic facts of this condition, which affects a little less than 1% of the population. They have severe reactions to the gluten in wheat. Celiac disease is mostly marked by gastroentereological complaints – diarrhea, bloating, abdominal pain – but it is also associated with vitamin deficiencies, anaemia, skin reactions, infertility, and "malaise". It can be pretty straightforwardly detected by blood tests and gut biopsies and is not subtle.

People start to disagree about the existence of "gluten sensitivity", which if it existed would be a bad reaction to gluten even in people who don't test positive for celiac disease. Many people believe they have gastrointestinal (or other) symptoms that go away when they eat gluten-free diets, but science can't find anything wrong with their intestines that could be causing the problems.

A recent study somewhat vindicated these people. Biesiekierski 2011 describes a double-blind randomized controlled trial: people who said they had "gluten-sensitive" irritable bowel syndrome were put on otherwise gluten-free diets and then randomly given either gluten or a placebo. They found that the patients given gluten reported symptoms (mostly bowel-related and tiredness) much more than those given placebo ($p = 0.0001$) but did not demonstrate any of the chemical, immunological, or histological markers usually associated with celiac disease. A similar Italian study found the same thing, except that they did find a higher rate of anti-gluten antibodies in their patients. Another study found that non-celiacs with antibodies to gluten had higher rates of mortality. And another study did find a histological change in bowel barrier function on this group of patients with the introduction of gluten. And another study from the same group found that maybe FODMAPs, another component of wheat, are equally or more responsible.

The journal *Gastroenterology*, which you may not be surprised to learn is the leading journal in the field of gastroenterology, proclaims:

> The current working definition of nonceliac gluten sensitivity (NCGS) is the occurrence of irritable bowel syndrome (IBS)-like symptoms after the ingestion of gluten and improvement after gluten withdrawal from the diet after exclusion of celiac disease based on negative celiac serologies and/or normal intestinal architecture and negative immunoglobulin (Ig)E-mediated allergy tests to wheat. Symptoms reported to be consistent with NCGS are both intestinal (diarrhea, abdominal discomfort or pain, bloating, and flatulence) and extra-intestinal (headache, lethargy, poor concentration, ataxia, or recurrent oral ulceration). These criteria strongly and conveniently suggest that NCGS is best understood as a subset of IBS or perhaps a closely related but distinct functional disorder. Although the existence of NCGS has been slowly gaining ground with physicians and scientists, NCGS has enjoyed rapid and widespread adoption by the general public.

But even this isn't really that interesting. Maybe some people with irritable bowel syndrome or certain positive antibodies should try avoiding gluten to see if it helps their specific and very real symptoms. At most ten percent of people are positive

antibody testing, and not all of those even have symptoms. That's still a far cry from saying no one should eat wheat.

But the anti-wheat crowd says an alternative more sensitive antibody test could raise sensitivity as high as a third of the population. The test seems to have been developed by a well-respected and legitimate doctor, but it hasn't as far as I can tell been submitted for peer review or been confirmed by any other source. Meh.

That's boring anyway. The real excitement comes from sweeping declarations that *the entire* population is sensitive to wheat.

Wheat Increases Intestinal Permeability Causing A Leaky Gut – *Probably true, of uncertain significance*

There are gluten-induced mucosal changes in subjects without small bowel disease. And gliadin increases intestinal permeability in the test tube, which should be extremely concerning to any test tubes reading this.

But probably the bigger worry here are lectins, which include wheat germ agglutinin. WGA affects the intestinal permeability of rats, which should be extremely concerning to any rats reading this. The same substance has been found to produce pro-inflammatory cytokines and interfere with the growth of various organs including the gut.

So there's pretty good evidence that chemicals in wheat can increase intestinal permeability. Who cares?

For years, "leaky gut syndrome" was an alternative medicine diagnosis that was soundly mocked by the mainstream medical establishment. Then the mainstream medical establishment confirmed it existed and did that thing where they totally excused their own mocking of it but were ABSOLUTELY OUTRAGED that the alternative medicine community might have in some cases been overenthusiastic about it.

Maybe I'm being too harsh. The alternative medicine community often does take "leaky gut syndrome" way too far.

On the other hand, it's probably real and *Nature Clinical Practice* is now publishing papers saying it is "a key ingredient in the pathogenesis of autoimmune diseases" and "offers innovative, unexplored approaches for the treatment of these devastating diseases" and gut health has been deemed "a new objective in medicine". Preliminary changes to intestinal permeability have been found in asthma, in diabetes, and even in depression.

But it's not yet clear if this is cause and effect. Maybe the stress of having asthma increases intestinal permeability somehow. Or maybe high intestinal permeability causes asthma somehow. It sure seems like the latter might work – all sorts of weird antigens and stuff from food can make it into the bloodstream and alarm the immune system – but right now this is all speculative.

So what we have is some preliminary evidence that wheat increases intestinal permeability, and some preliminary evidence that increased intestinal permeability is bad for you in a variety of ways.

And I don't doubt that those two facts are true, but my knowledge of this whole area is so weak that I wonder how much to worry.

What other foods increase intestinal permeability? Do they do it more or less than wheat? Has anyone been investigating this? Are there common things that affect intestinal permeability a thousand times more than wheat does, such that everything done by wheat is totally irrelevant in comparison?

Do people without autoimmune diseases suffer any danger from increased intestinal permeability? How much? Is it enough to offset the many known benefits of eating wheat (to be discussed later?) Fiber seems to decrease intestinal permeability and most people get their fiber from bread; would decreasing bread consumption make leaky gut even worse?

I find this topic really interesting, but in a "I hope they do more research" sort of way, not an "I shall never eat bread ever again" sort of way.

Digestion Of Wheat Produces Opiates, Which Get You Addicted To Wheat – *Probably false, but just true enough to be weird*

Dr. William Davis, a cardiologist, most famously makes this claim in his book *Wheat Belly*. He says that gliadin (a component of gluten) gets digested into opiates, chemicals similar to morphine and heroin with a variety of bioactive effects. This makes you addicted to food in general and wheat in particular, the same way you would get addicted to morphine or heroin. This is why people are getting fat nowadays – they're eating not because they're hungry, but because they're addicted. He notes that drugs that block opiates make people want wheat less.

Does Wheat Make Us Fat And Sick, a review published in the *Journal of Cereal Science* (they have journals for *everything* nowadays) is a good rebuttal to some of Davis' claims and a good pro-wheat resource in general.

They say that although gliadin does digest into opiates, those opiates are seven unit peptides and so too big to be absorbed from the gut to the bloodstream.

(note that having opiates *in your gut* isn't a great idea either since there are lots of nerves there controlling digestion that can be affected by these drugs)

But I'm not sure this statement about absorption is even true. First, large proteins can sometimes make it into the gut. Second, if all that leaky gut syndrome stuff above is right, maybe the gut is unusually permeable after wheat consumption. Third, there have been sporadically reported cases of gliadin-derived opiates found in the urine, which implied they got absorbed somehow.

There's a better counterargument on the blog The Curious Coconut. She notes that there's no evidence these peptides can cross the blood-brain barrier, a precondition for having any psychological effects. And although the opiate-blocker naloxone does decrease appetite, this effect is not preferential for wheat, and probably more related to the fact that opiates are the way the brain reminds itself it's enjoying itself (so that opiate-blocked people can't enjoy eating as much).

And then there's the usual absence of qualifiers. Lots of things are "chemically related" to other chemicals without having the same effect; are gliadin-derived opiates addictive? Are they produced in quantities high enough to be relevant in real life? Corn, spinach, and maybe meat can all get digested into opiates – is there any evidence wheat-derived opiates are worse? This is really sketchy.

The most convincing counterargument is that as far as anyone can tell, wheat makes people eat less, not more:

> Prospective studies suggest that weight gain and increases in abdominal adiposity over time are lower in people who consume more whole grains. Analyses of the Physicians' Health Study (27) and the Nurses' Health Study (26) showed that those who consumed more whole grain foods consistently weighed less than those who consumed fewer whole grain foods at each follow-up period of the study. Koh-Banerjee et al. (27) estimated that for every 40-g increase in daily whole grain intake, the 8-y weight gain was lower by 1.1 kg.

I'll discuss this in more detail later, but it does seem like a nail in the coffin for the "people eat too much because they're addicted to wheat" theory.

Still, who would have thought that wheat being digested into opiates was even a *little* true?

Wheat Something Something Something Autism And Schizophrenia – *Definitely weird*

Since gluten-free diets get tried for everything, and everything gets tried for autism, it was overdetermined that people would try gluten-free diets for autism.

All three of the issues mentioned above – immune reactivity to gluten, leaky guts, and gliadin-derived opiates – have been suggested as mechanisms for why gluten free diets might be useful in autism.

Of studies that have investigated, a review found that seven reported positive results, four negative results, and two mixed results – but that all of the studies involved were terrible and the ones that were slightly less terrible seemed to be more negative. The authors described this as evidence against gluten-free diets for autism, although someone with the opposite bias could have equally well looked at the same review and described it as supportive.

However, a very large epidemiological study found (popular article, study abstract) that people with antibodies to gluten had three times the incidence of autism spectrum disease than people without, and that the antibodies preceded the development of the condition.

Also, those wheat-derived opioids from the last section – as well as milk-derived opioids called casomorphins – seem to be detected at much higher rates in autistic people.

Both of these factors may have less to do with wheat in particular and more to do with some general dysregulation of peptide metabolism in autism. If for some reason the gut kept throwing peptides into the body inappropriately, this would disrupt neurodevelopment, lead to more peptides in the urine, and give the immune system more chance to react to gluten.

The most important thing to remember here is that it would be really wrong to say wheat might be "the cause" of autism. Most likely people do not improve on gluten-free diets. While there's room to argue that people might have picked up a small signal of them improving *a little*, the idea that this totally removes the condition is right out. If we were doing this same study with celiac disease, we wouldn't be

wasting our time with marginally significant results. Besides, we know autism is multifactorial, and we know it probably begins in utero.

Schizophrenia right now is in a similar place. Schizophrenics are five to seven times more likely to have anti-gliadin antibodies as the general population. We can come up with all sorts of weird confounders – maybe antipsychotic medications increase gut permeability? – but that's a really strong result. And schizophrenics have frank celiac disease at five to ten times the rate of the general population. Furthermore, a certain subset of schizophrenics sees a dramatic reduction in symptoms when put on a strict gluten-free diet (this is psychiatrically useless, both because we don't know which subset, and because given how much trouble we have getting schizophrenics to swallow one lousy pill every morning, the chance we can get them to stick to a gluten-free diet is basically nil). And like those with autism, schizophrenics show increased levels of weird peptides in their urine.

But a lot of patients with schizophrenia don't have reactions to gluten, a lot don't improve on a gluten free diet, and other studies question the research showing that any of them at all do.

The situation here looks a lot like autism – a complex multifactorial process that probably isn't caused by gluten but where we see interesting things going on in the vague territory of gluten/celiac/immune response/gut permeability/peptides, with goodness only knows which ones come first and which are causal.

Wheat Has Been Genetically Modified Recently In Ways That Make It Much Worse For You – *Probably true, especially if genetically modified means "not genetically modified" and "recently" means "nine thousand years ago"*

If you want to blame the "obesity epidemic" or "autism epidemic" or any other epidemic on wheat, at some point you have to deal with people eating wheat for nine thousand years and not getting epidemics of these things. Dr. Davis and other wheat opponents have turned to claims that wheat has been "genetically modified" in ways that improve crop yield but also make it more dangerous. Is this true?

Wheat has not been genetically modified in the classic sense, the one where mad scientists with a god complex inject genes from jellyfish into wheat and all of a sudden your bread has tentacles and every time you try to eat it it stings you. But it has been modified in the same way as all of our livestock, crops, and domestic pets – by selective breeding. Modern agricultural wheat doesn't look much like its ancient wild ancestors.

The *Journal Of Cereal Science* folk don't seem to think this is terribly relevant. They say:

> Gliadins are present in all wheat lines and in related wild species. In addition, seeds of certain ancient types of tetraploid wheat have even greater amounts of total gliadin than modern accessions…There is no evidence that selective breeding has resulted in detrimental effects on the nutritional properties or health benefits of the wheat grain, with the exception that the dilution of other components with starch occurs in modern high yielding lines (starch comprising about 80% of the grain dry weight). Selection for high protein content has been carried out for bread making, with modern bread making varieties generally containing about 1–2% more protein (on a grain dry weight basis) than varieties bred for livestock feed when grown under the same conditions. However, this genetically determined difference in protein content is less than can be achieved

> by application of nitrogen fertilizer. We consider that statements made in the book of Davis, as well as in related interviews, cannot be substantiated based on published scientific studies.

In support of this proposition, in the test tube ancient grains were just as bad for celiac patients' immune systems as modern ones.

And yet in one double-blind randomized-controlled trial, people with irritable bowel syndrome felt better on a diet of ancient grains than modern ones ($p < 0.0001$); and in another, people on an ancient grain diet had lower inflammatory markers and generally better nutritional parameters than people on a modern grain one. Isn't that interesting?

Even though it's a little bit weird and I don't think anyone understands the exact nutrients at work, sure, let's give this one to the ancient grain people.

The Lectins In Wheat Interfere With Leptin Receptors, Making People Leptin Resistant And Therefore Obese – *Currently at "mere assertion" level until I hear some evidence*

So here's the argument. Your brain has receptors for the hormone leptin, which tells you when to stop eating. But "lectin" sounds a lot like "leptin", and this confuses the receptors, so they give up and tell you to just eat as much as you want.

Okay, this probably isn't the real argument. But even though a lot of wheat opponents cite the heck out of this theory, the only presentation of evidence I can find is Jonsson et al (2005), which points out that there are a lot of diseases of civilization, they seem to revolve around leptin, something common to civilization must be causing them, and maybe that thing could be lectin.

But civilization actually contains more things than a certain class of proteins found in grains! There's poor evidence of lectin actually interfering with the leptin receptor in humans. The only piece of evidence they provide is a nonsignificant trend toward more cardiovascular disease in people who eat more whole grains in one study, and as we will see, that is wildly contradicted by all other studies.

This one does not impress me much.

Wheat Is Actually Super Good For You And You Should Have It All The Time – *Probably more evidence than the other claims on this list*

Before I mention any evidence, let me tell you what we're going to find.

We're going to find very, very many large studies finding conclusively that whole grains are great in a lot of different ways.

And we're not going to know whether it's at all applicable to the current question.

Pretty much all these studies show that people with some high level of "whole grain consumption" are much healthier than people with some lower level of same. That sounds impressive.

But what none of these studies are going to do a good job ruling out is that whole grain is just funging against refined grain which is even worse. Like maybe the people who report low whole grain consumption are eating lots of refined grain, and so more

total grain, and the high-whole-grain-consumption people are actually eating less grain total.

They're also not going to rule out the universal problem that if something is widely known to be healthy (like eating whole grains) then the same health-conscious people who exercise and eat lots of vegetables will start doing it, so when we find that the people doing it are healthier, for all we know it's just that the people doing it are exercising and eating vegetables.

That having been said, eating lots of whole grain decreases BMI, metabolic risk factors, fasting insulin, and body weight (1, 2, 3, 4,5.)

The American Society For Nutrition Symposium says:

> Several mechanisms have been suggested to explain why whole grain intake may play a role in body weight management. Fiber content of whole grain foods may influence food volume and energy density, gastric emptying, and glycemic response. Whole grains has also been proposed to play an important role in promoting satiety; individuals who eat more whole grain foods may eat less because they feel satisfied with less food. Some studies comparing feelings of fullness or actual food intake after ingestion of certain whole grains, such as barley, oats, buckwheat, or quinoa, compared with refined grain controls indicated a trend toward increased satiety with whole grains. These data are in accordance with analyses determining the satiety index of a large number of foods, which showed that the satiety index of traditional white bread was lower than that of whole grain breads. However, in general, these satiety studies have not observed a reduction in energy intake; hence, further research is needed to better understand the satiety effects of whole grains and their impact on weight management.
>
> Whole grains, in some studies, have also been observed to lower the glycemic and insulin responses, affect hunger hormones, and reduce subsequent food intake in adults. Ingestion of specific whole grains has been shown to influence hormones that affect appetite and fullness, such as ghrelin, peptide YY, glucose-dependent insulinotropic polypeptide, glucagon-like peptide 1, and cholecystokinin. Whole grain foods with fiber, such as wheat bran or functional doses of high molecular weight β-glucans, compared with lower fiber or refined counterparts have been observed to alter gastric emptying rates. Although it is likely that whole grains and dietary fiber may have similar effects on satiety, fullness, and energy intake, further research is needed to elucidate how, and to what degree, short-term satiety influences body weight in all age groups.
>
> Differences in particle size of whole grain foods may have an effect on satiety, glycemic response, and other metabolic and biochemical (leptin, insulin, etc.) responses. Additionally, whole grains have been suggested to have prebiotic effects. For example, the presence of oligosaccharides, RS, and other fermentable carbohydrates may increase the number of fecal bifidobacteria and lactobacilli (49), thus potentially increasing the SCFA production and thereby potentially altering the metabolic and physiological responses that affect body weight regulation.
>
> In summary, the current evidence among a predominantly Caucasian population suggests that consuming 3 or more servings of whole grains per day is associated with lower BMI, lower abdominal adiposity, and trends toward lower weight gain

> over time. However, intervention studies have been inconsistent regarding weight loss

The studies that combined whole and refined grains are notably fewer. But Dietary Intake Of Whole And Refined Grain Breakfast Cereals And Weight Gain In Men finds that among 18,000 male doctors, those who ate breakfast cereal (regardless of whether it was whole and refined) were less likely to become overweight several years later than those who did not ($p = 0.01$). A book with many international studies report several that find a health benefit of whole grains, several that find a health benefit of all grains (Swedes who ate more grains had lower abdominal obesity; Greeks who ate a grain-rich diet were less likely to become obese; Koreans who ate a "Westernized" bread-and-dairy diet were less likely to have abdominal obesity) and no studies that showed any positive association between grains and obesity, whether whole or refined.

I cannot find good interventional trials on what happens when a population replaces non-grain with grain.

On the other hand, Dr. Davis and his book *Wheat Belly* claim:

> Typically, people who say goodbye to wheat lose a pound a day for the first 10 days. Weight loss then slows to yield 25-30 pounds over the subsequent 3-6 months (differing depending on body size, quality of diet at the start, male vs. female, etc.)
>
> Recall that people who are wheat-free consume, on average, 400 calories less per day and are not driven by the 90-120 minute cycle of hunger that is common to wheat. It means you eat when you are hungry and you eat less. It means a breakfast of 3 eggs with green peppers and sundried tomatoes, olive oil, and mozzarella cheese for breakfast at 7 am and you're not hungry until 1 pm. That's an entirely different experience than the shredded wheat cereal in skim milk at 7 am, hungry for a snack at 9 am, hungry again at 11 am, counting the minutes until lunch. Eat lunch at noon, sleepy by 2 pm, etc. All of this goes away by banning wheat from the diet, provided the lost calories are replaced with real healthy foods."

Needless to say, he has no studies supporting this assertion. But the weird thing is, his message board is full of people who report having exactly this experience, my friends who have gone paleo have reported exactly this experience, and when I experimented with it, I had pretty much exactly this experience. Even the blogger from whom I took some of the strongest evidence criticizing Davis says she had exactly this experience.

The first and most likely explanation is that anecdotal evidence sucks and we should shut the hell up. Are there other, less satisfying explanations?

Maybe completely removing wheat from the diet has a nonlinear effect relative to cutting down on it? For example, in celiac disease there is no such thing as "partially gluten free" – if you have any gluten at all, your disease comes back in full force. This probably wouldn't explain Dr. Davis' observation – neither I nor my other wheatless-experimentation friends were as scrupulous as a celiac would have to be. But maybe there's a nonlinear discrepancy between people who have 75% the wheat of a normal person and 10% the wheat of a normal person?

Maybe there's an effect where people who like wheat but remove it from the diet are eating things they don't like, and so eat less of them? But people who don't like wheat

like other stuff, and so eat lots of that?

Maybe wheat in those studies is totally 100% a confounder for whether people are generally healthy and follow their doctor's advice, and the rest of the doctor's advice is really good but the wheat itself is terrible?

Maybe cutting out wheat has really positive short-term effects, but neutral to negative long-term effects?

Maybe as usual in these sorts of situations, the simplest explanation is best.

Final Thoughts

Non-celiac gluten sensitivity is clearly a real thing. It seems to produce irritable bowel type symptoms. If you have irritable bowel type symptoms, it might be worth trying a gluten-free diet for a while. But the excellent evidence for its existence doesn't seem to carry over to the normal population who don't experience bowel symptoms.

What these people have are vague strands of evidence. Something seems to be going on with autism and schizophrenia – but most people don't have autism or schizophrenia. The intestinal barrier seems to become more permeable with possible implications for autoimmune diseases – but most people don't have autoimmune disease. Some bad things seem to happen in rats and test tubes – but most people aren't rats or test tubes.

You'd have to want to take a position of maximum caution – wheat seems to do all these things, and even though none of them in particular obviously hurt me directly, all of them together make it look like the body just doesn't do very well with this substance, and probably other ways the body doesn't do very well with this substance will turn up, and some of them probably affect me.

There's honor in a position of maximum caution, especially in a field as confusing as nutrition. It would not surprise me if the leaky gut connection turned into something very big that had general implications for, for example, mental health. And then people who ate grain might regret it.

But stack that up against the pro-wheat studies. None of them are great, but they mostly do something the anti-wheat studies don't: show direct effect on things that are important to you. Most people don't have autism or schizophrenia, but most people *do* have to worry about cardiovascular disease. We *do* have medium-term data that wheat doesn't cause cancer, or increase obesity, or contribute to diabetes, or any of that stuff, and at this point solely based on the empirical data it seems much more likely to help with those things than hurt.

I hope the role of intestinal permeability in autoimmune disease gets the attention it deserves – and when it does, I might have to change my mind. I hope people stop being jerks about gluten sensitivity, admit it exists, and find better ways to deal with it. And if people find that eliminating bread from their diet makes them feel better or lose weight faster, cool.

But as far as I can tell the best evidence is on the pro-wheat side of things for most people at most times.

[**EDIT:** An especially good summary of the anti-wheat position is 6 Ways Wheat Can Destroy Your Health. An especially good pro-wheat summary is Does Wheat Make Us

Fat And Sick?]

SSRIs: Much More Than You Wanted To Know

The claim that “SSRIs don’t work” or “SSRIs are mostly just placebo” is most commonly associated with Irving Kirsch, a man with the awesome job title of “Associate Director Of The Program For Placebo Studies at Harvard”.

(fun fact: there’s actually no such thing as “Placebo Studies”, but Professor Kirsch’s *belief* that he directs a Harvard department inspires him to create much higher-quality research.)

In 1998, he published a meta-analysis of 19 placebo-controlled drug trials that suggested that almost all of the benefits of antidepressants were due to the placebo effect. Psychiatrists denounced him, saying that you can choose pretty much whatever studies you want for a meta-analysis.

After biding his time for a decade, in 2008 he struck back with another meta-analysis, this being one of the first papers in all of medical science to take the audacious step of demanding all the FDA’s data through the Freedom of Information Act. Since drug companies are required to report all their studies to the FDA, this theoretically provides a rare and wonderful publication-bias-free data set. Using this set, he found that, although antidepressants did seem to outperform placebo, the effect was not “clinically significant” except “at the upper end of very severe depression”.

This launched a minor war between supporters and detractors. Probably the strongest support he received was a big 2010 meta-analysis by Fournier et al, which found that

> The magnitude of benefit of antidepressant medication compared with placebo increases with severity of depression symptoms and may be minimal or nonexistent, on average, in patients with mild or moderate symptoms. For patients with very severe depression, the benefit of medications over placebo is substantial.

Of course, a very large number of antidepressants are given to people with mild or moderate depression. So what now?

Let me sort the debate about antidepressants into a series of complaints:

1. Antidepressants were oversold and painted as having more biochemical backing than was really justified
2. Modern SSRI antidepressants are no better than older tricyclic and MAOI antidepressants, but are prescribed much more because of said overselling
3. There is large publication bias in the antidepressant literature
4. The effect size of antidepressants is clinically insignificant
5. And it only becomes significant in the most severe depression
6. And even the effects found are only noticed by doctors, not the patients themselves
7. And even that unsatisfying effect might be a result of “active placebo” rather than successful treatment
8. And antidepressants have much worse side effects than you have been led to believe
9. Therefore, we should give up on antidepressants (except maybe in the sickest patients) and use psychotherapy instead

1. Antidepressants were oversold and painted as having more biochemical backing than was really justified – *Totally true*

It is starting to become slightly better known that the standard story – depression is a deficiency of serotonin, antidepressants restore serotonin and therefore make you well again – is kind of made up.

There was never much more evidence for the serotonin hypothesis than that chemicals that increased serotonin tended to treat depression – making the argument that "antidepressants are biochemically justified because they treat the low serotonin that is causing your depression" kind of circular. Saying "Serotonin treats depression, therefore depression is, at root, a serotonin deficiency" is about as scientifically grounded as saying "Playing with puppies makes depressed people feel better, therefore depression is, at root, a puppy deficiency".

The whole thing became less tenable with the discovery that several chemicals that didn't increase serotonin were also effective antidepressants – not to mention one chemical, tianeptine, that *decreases* serotonin. Now the conventional wisdom is that depression is a very complicated disturbance in several networks and systems within the brain, and serotonin is one of the inputs and/or outputs of those systems.

Likewise, a whole bunch of early '90s claims: that modern antidepressants have no side effects, that they produce miraculous improvements in everyone, that they make you better than well – seem kind of silly now. I don't think anyone is arguing against the proposition that there was an embarrassing amount of hype that has now been backed away from.

2. Modern SSRI antidepressants are no better than older tricyclic and MAOI antidepressants, but are prescribed much more because of said overselling – *First part true, second part less so*

Most studies find SSRI antidepressants to be no more effective in treating depression than older tricyclic and MAOI antidepressants. Most studies aren't really powered to do this. It seems clear that there aren't spectacular differences, and hunting for small differences has proven very hard.

If you're a geek about these sorts of things, you know that a few studies have found non-significant advantages for Prozac and Paxil over older drugs like clomipramine, and marginally-significant advantages for Effexor over SSRIs. But conventional wisdom is that tricyclics can be even more powerful than SSRIs for certain very severe hospitalized depression cases, and a lot of people think MAOIs worked better than anything out there today.

But none of this is very important because the real reason SSRIs are so popular is the side effect profile. While it is an exaggeration to say they have *no* side effects (see above) they are an obvious improvement over older classes of medication in this regard.

Tricyclics had a bad habit of causing fatal arrythmias when taken at high doses. This is really really bad in depression, because depressed people tend to attempt suicide and the most popular method of suicide attempt is overdosing on your pills. So if you give depressed people a pill that is highly fatal in overdose, you're basically enabling suicidality. This alone made the risk-benefit calculation for tricyclics unattractive in a lot of cases. Add in dry mouth, constipation, urinary problems, cognitive impairment, blurry vision, and the occasional tendency to cause heart arrythmias even when taken

correctly, and you have a drug you're not going to give people who just say they're feeling a little down.

MAOIs have their own problems. If you're using MAOIs and you eat cheese, beer, chocolate, beans, liver, yogurt, soy, kimchi, avocados, coconuts, et cetera, et cetera, et cetera, you have a chance of precipitating a "hypertensive crisis", which is exactly as fun as it sounds. As a result, people who are *already* miserable and *already* starving themselves are told they can't eat like half of food. And once again, if you tell people "Eat these foods with this drug and you die" and a week later the person wants to kill themselves and has some cheese in the house, then you're back to enabling suicide. There are some MAOIs that get around these restrictions in various clever ways, but they tend to be less effective.

SSRIs were the first class of antidepressants that mostly avoided these problems and so were pretty well-placed to launch a prescribing explosion even apart from being pushed by Big Pharma.

3. There is large publication bias in the antidepressant literature – *True, but not as important as some people think*

People became more aware of publication bias a couple of years after serious research into antidepressants started, and it's not surprising that these were a prime target. When this issue rose to scientific consciousness, several researchers tried to avoid the publication bias problem by using only FDA studies of antidepressants. The FDA mandates that its studies be pre-registered and the results reported no matter what they are. This provides a "control group" by which accusations of publication bias can be investigated. The results haven't been good. From Gibbons et al:

> Recent reports suggest that efficacy of antidepressant medications versus placebo may be overstated, due to publication bias and less efficacy for mildly depressed patients. For example, of 74 FDA-registered randomized controlled trials (RCTs) involving 12 antidepressants in 12,564 patients, 94% of published trials were positive whereas only 51% of all FDA registered studies were positive.

Turner et al express the same data a different way:

> . The FDA deemed 38 of the 74 studies (51%) positive, and all but 1 of the 38 were published. The remaining 36 studies (49%) were deemed to be either negative (24 studies) or questionable (12). Of these 36 studies, 3 were published as not positive, whereas the remaining 33 either were not published (22 studies) or were published, in our opinion, as positive (11) and therefore conflicted with the FDA's conclusion. Overall, the studies that the FDA judged as positive were approximately 12 times as likely to be published in a way that agreed with the FDA analysis as were studies with nonpositive results according to the FDA (risk ratio, 11.7; 95% confidence interval [CI], 6.2 to 22.0; P<0.001). This association of publication status with study outcome remained significant when we excluded questionable studies and when we examined publication status without regard to whether the published conclusions and the FDA conclusions were in agreement

The same source tells us about the effect this bias had on effect size:

> For each of the 12 drugs, the effect size derived from the journal articles exceeded the effect size derived from the FDA reviews (sign test, P<0.001). The magnitude of the increases in effect size between the FDA reviews and the published reports ranged from 11 to 69%, with a median increase of 32%. A 32% increase was also

> observed in the weighted mean effect size for all drugs combined, from 0.31 (95% CI, 0.27 to 0.35) to 0.41 (95% CI, 0.36 to 0.45).

I think a lot of this has since been taken on board, and most of the rest of the research I'll be talking about uses FDA data rather than published data. But as you can see, the overall change in effect size – from 0.31 to 0.41 – is not that terribly large.

4. The effect size of antidepressants is clinically insignificant – *Depends what you mean by "clinically insignificant"*

As mentioned above, when you try to control for publication bias, the effect size of antidepressant over placebo is 0.31.

This number can actually be broken down further. According to McAllister and Williams, who are working off of slightly different data and so get slightly different numbers, the effect size of placebo is 0.92 and the effect size of antidepressants is 1.24, which means antidepressants have a 0.32 SD benefit over placebo. Several different studies get similar numbers, including the Kirsch meta-analysis that started this whole debate.

Effect size is a hard statistic to work with (albeit *extremely fun*). The guy who invented effect size suggested that 0.2 be called "small", 0.5 be called "medium", and 0.8 be called "large". NICE, a UK health research group, somewhat randomly declared that effect sizes greater than 0.5 be called "clinically significant" and effect sizes less than 0.5 be called "not clinically significant", but their reasoning was basically that 0.5 was a nice round number, and a few years later they changed their mind and admitted they had no reason behind their decision.

Despite these somewhat haphazard standards, some people have decided that antidepressants' effect size of 0.3 means they are "clinically insignificant".

(please note that "clinically insignificant" is very different from "statistically insignificant" aka "has a p-value less than 0.05." Nearly everyone agrees antidepressants have a statistically significant effect – they do something. The dispute is over whether they have a clinically significant effect – the something they do is enough to make a real difference to real people)

There have been a couple of attempts to rescue antidepressants by raising the effect size. For example, Horder et al note that Kirsch incorrectly took the difference between the average effect of drugs and the average effect of placebos, rather than the average drug-placebo difference (did you follow that?) When you correct that mistake, the drug-placebo difference rises significantly to about 0.4.

They also note that Kirsch's study lumps all antidepressants together. This isn't necessarily wrong. But it isn't necessarily right, either. For example, his study used both Serzone (believed to be a weak antidepressant, rarely used) and Paxil (believed to be a stronger antidepressant, commonly used). And in fact, by his study, Paxil showed an effect size of 0.47, compared to Serzone's 0.21. But since the difference was not statistically significant, he averaged them together and said that "antidepressants are ineffective". In fact, his study showed that Paxil *was* effective, but when you average it together with a very ineffective drug, the effect disappears. He can get away with this because of the arcana of statistical significance, but by the same arcana I can get away with *not* doing that.

So right now we have three different effect sizes. 1.2 for placebo + drug, 0.5 for drug alone if we're being statistically merciful, 0.3 for drug alone if we're being harsh and letting the harshest critic of antidepressants pull out all his statistical tricks.

The reason effect size is *extremely fun* is that it allows you to compare effects in totally different domains. I will now attempt to do this in order to see if I can give you an intuitive appreciation for what it means for antidepressants.

Suppose antidepressants were in fact a weight loss pill.

An effect size of 1.2 is equivalent to the pill making you lose 32 lb.

An effect size of 0.5 is equivalent to the pill making you lose 14 lb.

An effect size of 0.3 is equivalent to the pill making you lose 8.5 lb.

Or suppose that antidepressants were a growth hormone pill taken by short people.

An effect size of 1.2 is equivalent to the pill making you grow 3.4 in.

An effect size of 0.5 is equivalent to the pill making you grow 1.4 in.

An effect size of 0.3 is equivalent to the pill making you grow 0.8 in.

Or suppose that antidepressants were a cognitive enhancer to boost IQ. This site gives us some context about occupations.

An effect size of 1.2 is equivalent to the pill making you gain 18 IQ points, ie from the average farm laborer to the average college professor.

An effect size of 0.5 is equivalent to the pill making you gain 7.5 IQ points, ie from the average farm laborer to the average elementary school teacher.

An effect size of 0.3 is equivalent to the pill making you gain 5 IQ points, ie from the average farm laborer to the average police officer.

To me, these kinds of comparisons are a little more revealing than NICE arbitrarily saying that anything below 0.5 doesn't count. If you could take a pill that helps your depression as much as gaining 1.4 inches would help a self-conscious short person, would you do it? I'd say it sounds pretty good.

5. The effect of antidepressants only becomes significant in the most severe depression – *Everything about this statement is terrible and everyone involved should feel bad*

So we've already found that saying antidepressants have an "insignificant" effect size is kind of arbitrary. But what about the second part of the claim – that they only have measurable effects in "the most severe depression"?

A lot of depression research uses a test called the HAM-D, which scores depression from 0 (none) to 52 (max). Kirsch found that the effect size of antidepressants increased as HAM-D scores increased, meaning antidepressants become more powerful as depression gets worse. He was only able to find a "clinically significant" effect size ($d > 0.5$) for people with HAM-D scores greater than 28. People have come up with various different mappings of HAM-D scores to words. For example, the APA says:

(0-7) No depression
(8-13) Mild depression
(14-18) Moderate depression
(19-22) Severe depression
(>=23) Very severe depression

Needless to say, a score of 28 sounds pretty bad.

We saw that Horder et al corrected some statistical deficiencies in Kirsch's original paper which made antidepressants improve slightly. With their methodology, antidepressants reach our arbitrary 0.5 threshold around HAM-D score 26. Another similar "antidepressants don't work" study got the number 25.

Needless to say, when anything over 23 is "very severe", 25 or 26 still sounds pretty bad.

Luckily, people completely disagree on the meanings of basic words! Very Severely Stupid is a cute article on Neuroskeptic that demonstrates that five different people and organizations suggest five different systems for rating HAM-D scores. Bech 1996 calls our 26 cutoff "major"; Funakawa 2007 calls it "moderate"; NICE 2009 calls it "severe". APA is unique in calling it very severe. NICE's scale is actually the exact same as the APA scale with every category renamed to sound one level less threatening. Facepalm.

Ghaemi and Vohringer(2011) go further and say that the real problem is that Kirsch is using the standard for depressive *symptoms*, but that real clinical practice involves depressive *episodes*. That is, all this "no depression" to "severe" stuff is about whether someone can be diagnosed with depression; presumably the people on antidepressants are definitely depressed and we need a new model of severity to determine just how depressed they are. As they put it:

> the authors of the meta-analysis claimed to use the American Psychiatric Association's criteria for severity of symptoms…in so doing, they ignore the obvious fact that symptoms differ from episodes: the typical major depressive episode (MDE) produced HDRS scores of at least 18 or above. Thus, by using symptom criteria, all MDEs are by definition severe or very severe. Clinicians know that some patients meet MDE criteria and are still able to work; indeed others frequently may not even recognize that such a person is clinically depressed. Other patients are so severe they function poorly at work so that others recognize something is wrong; some clinically depressed patients cannot work at all; and still others cannot even get out of bed for weeks or months on end. Clearly, there are gradations of severity within MDEs, and the entire debate in the above meta-analysis is about MDEs, not depressive symptoms, since all patients had to meet MDE criteria in all the studiesincluded in the meta-analysis (conducted by pharmaceutical companies for FDA approval for treatment of MDEs).
>
> The question, therefore, is not about severity of depressive symptoms, but severity of depressive episodes, assuming that someone meets DSM-IV criteria for a major depressive episode. On that question, a number of prior studies have examined the matter with the HDRS and with other depression rating scales, and the three groupings shown in table 2 correspond rather closely with validated and replicated definitions of mild (HDRS <24), moderate (HDRS 24–28), and severe (HDRS>28) major depressive episodes.

So, depending on whether we use APA criteria or G&V criteria, an HRDS of 23 is either "mild" (G&V) or "very severe" (APA).

Clear as mud? I agree that in one sense this is terrible. But in another sense it's actually a very important point. Kirsch's sample was really only "severe" in the context of everyone, both those who were clinically diagnosable with major depression and those who weren't. When we get to people really having a major depressive episode, a score of 26 to 28 isn't so stratospheric. But meanwhile:

> The APA seem to have ignored the fact that the HAMD did not statistically significantly distinguish between "Severe" and "Moderate" depression anyway (p=0.1)

Oh. That gives us some perspective, I guess. Also, some other people make the opposite critique and say that the HAM-D can't distinguish very well at the *low* end. Suppose HAM-Ds less than ten are meaningless and random. This would look a lot like antidepressants not working in mild depression.

Getting back to Ghaemi and Vohringer, they try a different tack and suggest that there is a statistical floor effect. They quite reasonably say that if someone had a HAM-D score of 30, and antidepressants solved 10% of their problem, they would lose 3 HAM-D points, which looks impressive. But if someone had a HAM-D score of 10, and antidepressants (still) solved 10% of their problem, they would only lose 1 HAM-D point, which sounds disappointing. But either way, the antidepressants are doing the same amount of work. If you adjust everything for baseline severity, it's easy to see that antidepressants here would have the same efficacy in severe and mild depression, even though it doesn't look that way at first.

I am confused that this works for effect sizes, because I expect effect sizes to be relative to the standard deviation in a sample. However, several important people tell me that it does, and that when you do this Kirsch's effect size goes from 0.32 to 0.40.

(I think these people are saying the exact same thing, but so overly mathematically that I've been staring at it for an hour and I'm still not certain)

More important, Ghaemi and Vohringer say once you do this, antidepressants reach the magic 0.5 number not only in severe depression, but also in moderate depression. However, when I look at this claim closely, almost all the work is done by G&V's adjusted scale in which Kirsch's "very severe" corresponds to their "mild".

(personal aside: I got an opportunity to talk to Dr. Ghaemi about this paper and clear up some of my confusion. Well, not exactly an opportunity to talk about it, per se. Actually, he was supposed to be giving me a job interview at the time. I guess we both got distracted. This may be one of several reasons I do not currently work at Tufts.)

So. In conclusion, everyone has mapped HAM-D numbers into words like "moderate" in totally contradictory ways, such that one person's "mild" is another person's "very severe". Another person randomly decided that we can only call things "clinically significant" if they go above the nice round number of 0.5, then retracted this. So when people say "the effects of antidepressants are only clinically significant in severe depression", what they mean is "the effects of antidepressants only reach a totally arbitrary number one guy made up and then retracted, in people whose HAM-D score is above whatever number I make up right now." Depending on what number you choose and what word you make up to describe it, you can find that

antidepressants are useful in moderate depression, or severe depression, or super-duper double-dog-severe depression, or whatever.

Science!

6. The beneficial effects of antidepressants are only noticed by doctors, not the patients themselves – *Partly true but okay*

So your HAM-D score has gone down and you're no longer officially in super-duper double-dog severe depression anymore. What does that mean for the patient?

There are consistent gripes that antidepressant studies that use patients rating their own mood show less improvement than studies where doctors rate how they think a patient is doing, or standardized tests like the HAM-D.

Some people try to turn this into a conspiracy, where doctors who have somehow broken the double-blinding of studies try to report that patients have done better because doctors like medications and want them to succeed.

The reality is more prosaic. It has been known for forty years that people's feelings are the last thing to improve during recovery from depression.

This might sound weird – what is depression except people's feelings? But the answer is "quite a lot". Depressed people often eat less, sleep more, have less energy, and of course are more likely to attempt suicide. If a patient gets treated with an antidepressant, and they start smiling more and talking more and getting out of the house and are no longer thinking about suicide, their doctor might notice – but the patient herself might still feel really down-in-the-dumps.

I am going to get angry comments from people saying I am declaring psychiatric patients too stupid to notice their own recovery or something like that, but it is a very commonly observed phenomenon. Patients have access to internal feelings which they tend to weight much more heavily than external factors like how much they are able to get done during a day or how many crying spells they have, sometimes so much so that they completely miss these factors. Doctors (or family members, or other outside observers) who don't see these internal feelings, are better able to notice outward signs. As a result, it is pretty universally believed that doctors spot signs of recovery in patients long before the patients themselves think they are recovering. This isn't just imaginary – it's found it datasets where the doctors are presumably blinded and with good inter-rater reliability.

Because most antidepressant trials are short, a lot of them reach the point where doctors notice improvement but not the point where patients notice quite as much improvement.

7. The apparent benefits of antidepressant over placebo may be an "active placebo" effect rather than a drug effect – *Unlikely*

Active placebo is the uncomfortable idea that no study can really have a blind control group because of side effects. That is, sugar pills have no side effects, real drugs generally do, and we all know side effects are how you know that a drug is working!

(there is a counterargument that placebos very often have placebo side effects, but most likely the real drug will at least have *more* side effects, saving the argument)

The solution is to use active placebo, a drug that has side effects but, as far as anyone knows, doesn't treat the experimental condition (in this case, depression). The preliminary results from this sort of study don't look good for antidepressants:

> Thomson reviewed 68 double-blind studies of tricyclics that used an inert placebo and seven that used an active placebo (44). He found drug efficacy was demonstrated in 59% of studies that employed inert placebo, but only 14% of those that used active placebo (?2=5.08, df=1, p=0.02). This appears to demonstrate that in the presence of a side-effect-inducing control condition, placebo cannot be discriminated from drug, thus affirming the null hypothesis.

Luckily, Quitkin et al (2000) solve this problem so we don't have to:

> Does the use of active placebo increase the placebo response rate? This is not the case. After pooling data from those studies in which a judgment could be made about the proportion of responders, it was found that 22% of patients (N=69 of 308) given active placebos were rated as responders. To adopt a conservative stance, one outlier study (50) with a low placebo response rate of 7% (N=6 of 90) was eliminated because its placebo response rate was unusually low (typical placebo response rates in studies of depressed outpatients are 25%–35%). Even after removing this possibly aberrant placebo group, the aggregate response rate was 29% (N=63 of 218), typical of an inactive placebo. The active placebo theory gains no support from these data.
>
> Closer scrutiny suggests that the "failure" of these 10 early studies to find typical drug-placebo differences is attributable to design errors that characterize studies done during psychopharmacology's infancy. Eight of the 10 studies had at least one of four types of methodological weaknesses: inadequate sample size, inadequate dose, inadequate duration, and diagnostic heterogeneity. The flaws in medication prescription that characterize these studies are outlined in Table 3. In fact, in spite of design measurement and power problems, six of these 10 studies still suggested that antidepressants are more effective than active placebo.
>
> In summary, these reviews failed to note that the active placebo response rate fell easily within the rate observed for inactive placebo, and the reviewers relied on pioneer studies, the historical context of which limits them.

In other words, active placebo research has fallen out of favor in the modern world. Most studies that used active placebo are very old studies that were not very well conducted. Those studies failed to find an active-placebo-vs.-drug difference because they weren't good enough to do this. But they also failed to find an active-placebo-vs.-inactive-placebo difference. So they provide no support for the idea that active placebos are stronger than inactive placebos in depression and in fact somewhat weigh against it.

8. Antidepressants have much worse side effects than you were led to believe – *Depends how bad you were led to believe the side effects were*

As discussed in Part 2, the biggest advantage of SSRIs and other new antidepressants over the old antidepressants was their decreased side effect profile. This seems to be quite real. For example, Brambilla finds a relative risk of adverse events on SSRIs only 60% of that on TCAs, p = 0.003 (although there are some conflicting numbers in that paper I'm not really clear about). Montgomery et al 1994 finds that fewer patients stop taking SSRIs than tricyclics (usually a good "revealed preference"-style measure of side effects since sufficiently bad side effects make you stop using the drug).

The charmingly named Cascade, Kalali, and Kennedy (2009) investigated side effect frequency in a set of 700 patients on SSRIs and found the following:

56% decreased sexual functioning
53% drowsiness
49% weight gain
19% dry mouth
16% insomnia
14% fatigue
14% nausea
13% light-headedness
12% tremor

However, it is very important to note that this study was not placebo controlled. Placebos can cause *terrible* side effects. Anybody who experiments with nootropics know that the average totally-useless inactive nootropic causes you to suddenly imagine all sorts of horrible things going on with your body, or attribute some of the things that happen anyway ("I'm tired") to the effects of the pill. It's not really clear how much of the stuff in this study is placebo effect versus drug effect.

Nevertheless, it is worth mentioning that 34% of patients declare side effects "not at all" or "a litte" bothersome, 40% "somewhat" bothersome, and 26% "very" or "extremely" bothersome. That's much worse than I would have expected.

Aside from the sort of side effects that you expect with any drug, there are three side effects of SSRIs that I consider especially worrisome and worthy of further discussion. These are weight gain, sexual side effects, and emotional blunting.

Weight gain is often listed as one of the most common and debilitating effects of SSRIs. But amusingly, when a placebo-controlled double-blinded study was finally run, SSRIs produced less weight gain than placebo. After a year of pill-taking, people on Prozac had gained 3.1 kg; people on placebo had gained 4.3. There is now some talk of SSRIs as a weak but statistically significant agent for weight *loss*.

What happened? One symptom of depression is not eating. People get put on SSRIs when they're really depressed. Then they get better, either because the drugs worked, because of placebo, or just out of regression to the mean. When you go from not eating to eating, you gain weight. In the one-year study, almost everyone's depression remitted (even untreated depressive episodes rarely last a whole year), so everyone went from a disease that makes them eat less, to remission from that disease, so everyone gained weight.

Sexual side effects are a less sanguine story. Here the direction was opposite: the medical community went from thinking this was a minor problem to finding it near-universal. The problem was that doctors usually just ask "any side effects?", and off Tumblr people generally don't volunteer information about their penis or vagina to a stranger. When they switched to the closed-ended question "Are you having any sexual side effects?", a lot of people who denied side effects in general suddenly started talking.

Numbers I have heard for the percent of people on SSRIs with sexual side effects include 14, 24, 37, 58, 59, and 70 (several of those come from here. After having read quite a bit of this research, I suspect you've got at least a 50-50 chance (they say men are more likely to get them, but they're worse in women). Of people who develop

sexual side effects, 40% say they caused serious distress, 35% some distress, and 25% no distress.

So I think it is fair to say that if you are sexually active, your chances with SSRIs are not great. Researchers investigating the topic suggest people worried about sexual side effects should switch to alternative sexual-side-effect-free antidepressant Serzone. You may remember that as the antidepressant that worked worst in the efficacy studies and brought the efficacy of all the other ones down with it. Also, it causes liver damage. In my opinion, a better choice would be bupropion, another antidepressant which has been found many times not to cause sexual side effects and which may even improve your sex life.

("Bupropion lacks this side effect" is going to be a common theme throughout this section. Bupropion causes insomnia, decreased appetite, and in certain rare cases of populations at risk, seizures. It is generally a good choice for people who are worried about SSRI side effects and would prefer a totally *different* set of side effects.)

There is a certain feeling that, okay, these drugs may have very very common, possibly-majority-of-user sexual side effects, but depressed people probably aren't screwing like rabbits anyway. So after you recover, you can wait the appropriate amount of time, come off the drugs (or switch to a different drug or dose for maintenance) and no harm done.

The situation no longer seems so innocuous. Despite a lack of systematic investigation, there are multiple reports from researchers and clinicians – not to mention random people on the Internet – of permanent SSRI-induced sexual dysfunction that does not remit once the drug is stopped. This is *definitely* not the norm and as far as we know it is so rare as to be unstudyable beyond the occasional case report.

On the other hand, I have this. I took SSRIs for about five to ten years as a kid, and now I have approximately the pattern of sexual dysfunction associated with SSRIs and consider myself asexual. Because I started the SSRIs too early to observe my sexuality without them, I can't officially blame the drugs. But I am very suspicious. I feel like this provides moderate anthropic evidence that it is not as rare as everyone thinks.

The last side effect worth looking at is emotional blunting. A lot of people say they have trouble feeling intense emotions (sometimes: any emotions at all) when on SSRIs. Sansone and Sansone (2010) report:

> As for prevalence rates, according to a study by Bolling and Kohlenberg, approximately 20 percent of 161 patients who were prescribed an SSRI reported apathy and 16.1 percent described a loss of ambition. In a study by Fava et al, which consisted of participants in both the United States and Italy, nearly one-third on any antidepressant reported apathy, with 7.7 percent describing moderate-to-severe impairment, and nearly 40 percent acknowledged the loss of motivation, with 12.0 percent describing moderate-to-severe impairment.

A practicing clinician working off observation finds about the same numbers:

> The sort of emotional "flattening" I have described with SSRIs may occur, in my experience, in perhaps 10-20% of patients who take these medications…I do want to emphasize that most patients who take antidepressant medication under careful medical supervision do not wind up feeling "flat" or unable to experience

> life's normal ups and downs. Rather, they find that–in contrast to their periods of severe depression–they are able to enjoy life again, with all its joys and sorrows.

Many patients who experience this side effect note that when you're depressed, "experiencing all your emotions fully and intensely" is not very high on your list of priorities, since your emotions tend to be terrible. There is a subgroup of depressed patients whose depression takes the form of not being able to feel anything at all, and I worry this effect would exacerbate their problem, but I have never heard this from anyone and SSRIs do not seem less effective in that subgroup, so these might be two different things that only sound alike. A couple of people discussing this issue have talked about how decreased emotions help them navigate interpersonal relationships that otherwise might involve angry fights or horrible loss – which sounds plausible but also really sad.

According to Barnhart et al (2004), "this adverse effect has been noted to be dose-dependent and reversible" – in other words, it will get better if you cut your dose, and go away completely when you stop taking the medication. I have not been able to find any case studies or testimonials by people who say this effect has been permanent.

My own experience was that I did notice this (even before I knew it was an official side effect) that it did go away after a while when I stopped the medications, and that since my period of antidepressant use corresponded with an important period of childhood socialization I ended out completely unprepared for having normal emotions and having to do a delicate social balancing act while I figured out how to cope with them. Your results may vary.

There is also a large research on suicidality as a potential side effect of SSRIs, but this looks like it would require another ten thousand words just on its own, so let's agree it's a risk and leave it for another day.

9. Therefore, we should give up on medication and use psychotherapy instead – *Makes sense right up until you run placebo-controlled trials of psychotherapy*

The conclusion of these studies that claim antidepressants don't outperform placebo is usually that we should repudiate Big Pharma, toss the pills, and go back to using psychotherapy.

The implication is that doctors use pills because they think they're much more effective than therapy. But that's not really true. The conventional wisdom in psychiatry is that antidepressants and psychotherapy are about equally effective.

SSRIs get used more than psychotherapy for the same reason they get used more than tricyclics and MAOIs – not because they're better but because they have fewer problems. The problem with psychotherapy is you've got to get severely mentally ill people to go to a place and talk to a person several times a week. Depressed people are not generally known for their boundless enthusiasm for performing difficult tasks consistently. Also, Prozac costs like 50 cents a pill. Guess how much an hour of a highly educated professional's time costs? More than 50c, that's for sure. If they are about equal in effectiveness, you probably don't want to pay extra and your insurance *definitely* doesn't want to pay extra.

Contrary to popular wisdom, it is almost never the doctor pushing pills on a patient who would prefer therapy. If anything it's more likely to be the opposite.

However, given that we're acknowledging antidepressants have an effect size of only about 0.3 to 0.5, is it time to give psychotherapy a second look?

No. Using very similar methodology, a team involving Mind The Brain blogger James Coyne found that psychotherapy decreases HAM-D scores by about 2.66, very similar to the 2.7 number obtained by re-analysis of Kirsch's data on antidepressants. It concludes:

> Although there are differences between the role of placebo in psychotherapy and pharmacotherapy research, psychotherapy has an effect size that is comparable to that of antidepressant medications. Whether these effects should be deemed clinically relevant remains open to debate.

Another study by the same team finds psychotherapy has an effect size of 0.22 compared to antidepressants' 0.3 – 0.5, though no one has tried to check if that difference is statistically significant and this does not give you the right to say antidepressants have "outperformed" psychotherapy.

If a patient has the time, money, and motivation for psychotherapy, it may be a good option – though I would only be comfortable using it as a monotherapy if the depression was relatively mild.

10. Further complications

What if the small but positive effect size of antidepressants wasn't because they had small positive effects on everyone, but because they had very large positive effects on some people, and negative effects on others, such that it averaged out to small positive effects? This could explain the clinical observations of psychiatrists (that patients seem to do *much* better on antidepressants) without throwing away the findings of researchers (that antidepressants have only small benefits over placebo) by bringing in the corollary that some psychiatrists notice some patients doing poorly on antidepressants and stop them in those patients (which researchers of course would not do).

This is the claim of Gueorguieva and Krystal 2011, who used "growth modeling" to analyze seven studies of new-generation-antidepressant Cymbalta and found statistically significant differences between two "trajectories" for the drug, but not for placebo. 66% of people were in the "responder" trajectory and outperformed placebo by 6 HAM-D points (remember, previous studies estimated HAM-D benefits over placebo at about 2.7). 33% of people were nonresponders and did about 6 HAM-D points *worse* than placebo. Average it out, and people did about 3 HAM-D points better on drug and placebo, pretty close to the previous 2.7 point estimate.

I don't know enough about growth modeling to be sure that the researchers didn't just divide the subjects into two groups based on treatment efficacy and say "Look! The subsection of the population whom we selected for doing well did well!" but they use many complicated statistics words throughout the study that I think are supposed to indicate they're not doing this.

If true, this is very promising. It means psychiatrists who are smart enough to notice people getting worse on antidepressants can take them off (or switch to another class of medication) and expect the remainder to get much, *much* better. I await further research with this methodology.

What if there were actually no such thing as the placebo effect? I know dropping this in around the end of an essay that assumes 75% of gains related to antidepressants are due to the placebo effect is a bit jarring, but it is the very-hard-to-escape conclusion of Hróbjartsson and Gøtzsche's meta-analysis on placebo. They find that three-armed studies – ie those that have a no-treatment group, a placebo-treatment group, and a real-drug-treatment group – rarely find much of a difference between no-treatment and placebo. This was challenged by Wampold et al here and here, but defended against those challenges by the long-name-Scandinavian-people here. Kirsch, who between all his antidepressant work is *still* Associate Director of Placebo Studies, finds here that 75% of the apparent placebo effect in antidepressant studies is probably a real placebo effect, but his methodology is a valiant attempt to make the most out of a total lack of data rather than a properly-directed study per se.

If placebo pills don't do much, what explains the vast improvements seen in both placebo and treatment groups in antidepressant trials? It could be the feeling of cared-for-ness and special-ness of getting to see a psychiatrist and talk with her about your problems, and the feeling of getting-to-contribute-something you get from participating in a scientific study. Or it could just be regression to the mean – most people start taking drugs when they feel very depressed, and at some point you have nowhere to go but up. Most depression gets better after six months or so – which is a much longer period than the six week length of the average drug trial, but maybe some people only volunteered for the study four months and two weeks after their depression started.

If Hróbjartsson and Gøtzsche were right, and Kirsch and the psychiatric establishment wrong, what would be the implications? Well, the good implication is that we no longer have to worry about problem 7 – that antidepressants are merely an active placebo – since active placebos shouldn't do anything. That means we can be more confident they really work. The more complicated implication is that psychiatrists lose one excuse for asking people to take the drugs – "Sure, the drug effect may be small, but the placebo effect is so strong that it's still worth it." I don't know how many psychiatrists actually think this way, but *I* sometimes think this way.

What if the reason people have so much trouble finding good effects from antidepressants is that they're giving the medications wrong? Psychiatric Times points out that:

> The Kirsch meta-analysis looked only at studies carried out before 1999. The much-publicized Fournier study examined a total of 6 antidepressant trials (n=718) using just 2 antidepressants, paroxetine and imipramine. Two of the imipramine studies used doses that were either subtherapeutic (100 mg/day) or less than optimal (100 to 200 mg/day)

What if we've forgotten the most important part? Antidepressants are used not only to treat acute episodes of depression, but to prevent them from coming back (maintenance therapy). This they apparently do very well, and I have seen very few studies that attempt to call this effect into question. Although it is always possible that someone will find the same kind of ambiguity around maintenance antidepressant treatment as now clouds acute antidepressant treatment, so far as far as I know this has not happened.

What if we don't understand what's going on with the placebo effect in our studies? Placebo effect has consistently gotten stronger over the past few decades, such that the difference between certain early tricyclic studies (which often found strong

advantages for the medication) and modern SSRI studies (which often find only weak advantages for the medication) is not weaker medication effect, but stronger placebo effect (that is, if medication always has an effect of 10, but placebo goes from 0 to 9, apparent drug-placebo difference gets much lower). Wired has a good article on this. Theories range from the good – drug company advertising and increasing prestige and awareness of psychiatry have raised people's expectations of psychiatric drugs – to the bad – increasing scientific competence and awareness have improved blinding and other facets of trial design – to the ugly – modern studies recruit paid participants with advertisements, so some unscrupulous people may be entering studies and then claiming to get better, hoping that this sounds sufficiently like the outcome the researchers want that everyone will be happy and they'll get their money on schedule.

If placebos are genuinely getting better because of raised expectations, that's good news for doctors and patients but bad news for researchers and drug companies. The patient will be happy because they get better no matter how terrible a prescribing decision the doctor makes; the doctor will be happy because they get credit. But for researchers and drug companies, it means it's harder to prove a difference between drug and placebo in a study. You can invent an excellent new drug and still have it fail to outperform placebo by very much if everyone in the placebo group improves dramatically.

Conclusion

An important point I want to start the conclusion section with: no matter what else you believe, antidepressants are not literally ineffective. Even the most critical study – Kirsch 2008 – finds antidepressants to outperform placebo with $p < .0001$ significance.

An equally important point: everyone except those two Scandinavian guys with the long names agree that, if you count the placebo effect, antidepressants are extremely impressive. The difference between a person who gets an antidepressant and a person who gets no treatment at all is like night and day.

The debate takes place within the bounds set by those two statements.

Antidepressants give a very modest benefit over placebo. Whether this benefit is so modest as to not be worth talking about depends on what level of benefits you consider so modest as to not be worth talking about. If you are as depressed as the average person who participates in studies of antidepressants, you can expect an antidepressant to have an over-placebo-benefit with an effect size of 0.3 to 0.5. That's the equivalent of a diet pill that gives you an average weight loss of 9 to 14 pounds, or a growth hormone that makes you grow on average 0.8 to 1.4 inches.

You may be able to get more than that if you focus on the antidepressants, like paroxetine and venlafaxine, that perform best in studies, but we don't have the statistical power to say that officially. It may be the case that most people who get antidepressants do much better than that but a few people who have paradoxical negative responses bring down the average, but right now this result has not been replicated.

This sounds moderately helpful and probably well worth it if the pills are cheap (which generic versions almost always are) and you are not worried about side effects. Unfortunately, SSRIs do have some serious side effects. Some of the supposed side effects, like weight gain, seem to be mostly mythical. Others, like sexual dysfunction, seem to be very common and legitimately very worrying. You can avoid most of these

side effects by taking other antidepressants like bupropion, but even these are not totally side-effect free.

Overall I think antidepressants come out of this definitely not looking like perfectly safe miracle drugs, but as a reasonable option for many people with moderate (aka "mild", aka "extremely super severe") depression, especially if they understand the side effects and prepare for them.

Alcoholics Anonymous: Much More Than You Wanted To Know

[EDIT 10/27: Slight changes in response to feedback; correcting some definitions. I am not an expert in this field and will continue to make changes as I learn about them. There is a critique of this post here and other worse critiques elsewhere. My only excuse for doing this is that I am failing less spectacularly than other online sources writing about the same topic.]

I've worked with doctors who think Alcoholics Anonymous is so important for the treatment of alcoholism that anyone who refuses to go at least three times a week is in denial about their problem and can't benefit from further treatment.

I've also worked with doctors who are so against the organization that they describe it as a "cult" and say that a physician who recommends it is no better than one who recommends crystal healing or dianetics.

I finally got so exasperated that I put on my Research Cap and started looking through the evidence base.

My conclusion, after several hours of study, is that now I understand why most people don't do this.

The studies surrounding Alcoholics Anonymous are some of the most convoluted, hilariously screwed-up research I have ever seen. They go wrong in ways I didn't even realize research *could* go wrong before. Just to give some examples:

– In several studies, subjects in the "not attending Alcoholics Anonymous" condition attended Alcoholics Anonymous more than subjects in the "attending Alcoholics Anonymous" condition.

– Almost everyone's belief about AA's retention rate is off by a factor of five because one person long ago misread a really confusing graph and everyone else copied them without double-checking.

– The largest study ever in the field, a $30 million effort over 8 years following thousands of patients, had no untreated control group.

Not only are the studies poor, but the people interpreting them are heavily politicized. The entire field of addiction medicine has gotten stuck in the middle of some of the most divisive issues in our culture, like whether addiction is a biological disease or a failure of willpower, whether problems should be solved by community and peer groups or by highly trained professionals, and whether there's a role for appealing to a higher power in any public organization. AA's supporters see it as a scruffy grassroots organization of real people willing to get their hands dirty, who can cure addicts failed time and time again by a system of glitzy rehabs run by arrogant doctors who think their medical degrees make them better than people who have personally fought their own battles. Opponents see it as this awful cult that doesn't provide any real treatment and just tells addicts that they're terrible people who will never get better unless they sacrifice their identity to the collective.

As a result, the few sparks of light the research kindles are ignored, taken out of context, or misinterpreted.

The entire situation is complicated by a bigger question. We will soon find that AA usually does not work better or worse than various other substance abuse interventions. That leaves the sort of question that all those fancy-shmancy people with control groups in their studies don't have to worry about – does anything work at all?

I.

We can start by just taking a big survey of people in Alcoholics Anonymous and seeing how they're doing. On the one hand, we don't have a control group. On the other hand…well, there really is no other hand, but people keep doing it.

According to AA's own surveys, one-third of new members drop out by the end of their first month, half by the end of their third month, and three-quarters by the end of their first year. "Drop out" means they don't go to AA meetings anymore, which could be for any reason including (if we're feeling optimistic) them being so completely cured they no longer feel they need it.

There is an alternate reference going around that only 5% (rather than 25%) of AA members remain after their first year. This is a mistake caused by misinterpreting a graph showing that only five percent of members in their first year were in their twelfth month of membership, which is obviously completely different. Nevertheless, a large number of AA hate sites (and large rehabs!) cite the incorrect interpretation, for example the Orange Papers and RationalWiki's page on Alcoholics Anonymous. In fact, just to keep things short, assume RationalWiki's AA page makes every single mistake I warn against in the rest of this article, then use that to judge them in general. On the other hand, Wikipedia gets it right and I continue to encourage everyone to use it as one of the most reliable sources of medical information available to the public (I wish I was joking).

This retention information isn't very helpful, since people can remain in AA without successfully quitting drinking, and people may successfully quit drinking without being in AA. However, various different sources suggest that, of people who stay in AA a reasonable amount of time, about half stop being alcoholic. These numbers can change wildly depending on how you define "reasonable amount of time" and "stop being alcoholic". Here is a table, which I have cited on this blog before and will probably cite again:

Behold. Treatments that look very impressive (80% improved after six months!) turn out to be the same or worse as the control group. And comparing control group to control group, you can find that "no treatment" can appear to give wildly different outcomes (from 20% to 80% "recovery") depending on what population you're looking at and how you define "recovery".

Twenty years ago, it was extremely edgy and taboo for a reputable scientist to claim that alcoholics could recover on their own. This has given way to the current status

quo, in which pretty much everyone in the field writes journal articles all the time about how alcoholics can recover on their own, but make sure to harp upon how edgy and taboo they are for doing so. From these sorts of articles, we learn that about 80% of recovered alcoholics have gotten better without treatment, and many of them are currently able to drink moderately without immediately relapsing (something *else* it used to be extremely taboo to mention). Kate recently shared an good article about this: Most People With Addiction Simply Grow Out Of It: Why Is This Widely Denied?

Anyway, all this stuff about not being able to compare different populations, and the possibility of spontaneous recovery, just mean that we need controlled experiments. The largest number of these take a group of alcoholics, follow them closely, and then evaluate all of them – the AA-attending and the non-AA-attending – according to the same criteria. For example Morgenstern et al (1997), Humphreys et al (1997) and Moos (2006). Emrick et al (1993) is a meta-analyses of *a hundred seventy three* of these. All of these find that the alcoholics who end up going to AA meetings are much more likely to get better than those who don't. So that's good evidence the group is effective, right?

Bzzzt! No! Wrong! Selection bias!

People who want to quit drinking are more likely to go to AA than people who don't want to quit drinking. People who want to quit drinking are more likely to *actually* quit drinking than those who don't want to. This is a *serious* problem. Imagine if it is common wisdom that AA is the best, maybe the only, way to quit drinking. Then 100% of people who really want to quit would attend compared to 0% of people who didn't want to quit. And suppose everyone who wants to quit succeeds, because secretly, quitting alcohol is really easy. Then 100% of AA members would quit, compared to 0% of non-members – the most striking result it is mathematically possible to have. And yet AA would not have made a smidgeon of difference.

But it's worse than this, because attending AA isn't just about wanting to quit. It's also about having the resources to make it to AA. That is, wealthier people are more likely to hear about AA (better information networks, more likely to go to doctor or counselor who can recommend) and more likely to be able to attend AA (better access to transportation, more flexible job schedules). But wealthier people are also known to be better at quitting alcohol than poor people – either because the same positive personal qualities that helped them achieve success elsewhere help them in this battle as well, or just because they have fewer other stressors going on in their lives driving them to drink.

Finally, perseverance is a confounder. To go to AA, and to keep going for months and months, means you've got the willpower to drag yourself off the couch to do a potentially unpleasant thing. That's probably the same willpower that helps you stay away from the bar.

And then there's a confounder going the *opposite* direction. The worse your alcoholism is, the more likely you are to, as the organization itself puts it, "admit you have a problem".

These sorts of longitudinal studies are almost useless and the field has mostly moved away from them. Nevertheless, if you look on the pro-AA sites, you will find them in droves, and all of them "prove" the organization's effectiveness.

III.

It looks like we need randomized controlled trials. And we have them. Sort of.

Brandsma (1980)

is the study beloved of the AA hate groups, since it purports to show that people in Alcoholics Anonymous not only don't get better, but are *nine times* more likely to binge drink than people who don't go into AA at all.

There are a number of problems with this conclusion. First of all, if you actually look at the study, this is one of about fifty different findings. The other findings are things like "88% of treated subjects reported a reduction in drinking, compared to 50% of the untreated control group".

Second of all, the increased binge drinking was significant at the 6 month followup period. It was *not* significant at the end of treatment, the 3 month followup period, the 9 month followup period, or the 12 month followup period. Remember, taking a single followup result out of the context of the other followup results is a classic piece of Dark Side Statistics and will send you to Science Hell.

Of multiple different endpoints, Alcoholics Anonymous did better than no treatment on almost all of them. It did worse than other treatments on some of them (dropout rates, binge drinking, MMPI scale) and the same as other treatments on others (abstinent days, total abstinence).

If you are pro-AA, you can say "Brandsma study proves AA works!". If you are anti-AA, you can say "Brandsma study proves AA works worse than other treatments!", although in practice most of these people prefer to quote extremely selective endpoints out of context.

However, most of the patients in the Brandsma study were people convicted of alcohol-related crimes ordered to attend treatment as part of their sentence. Advocates of AA make a good point that this population might be a bad fit for AA. They may not feel any personal motivation to treatment, which might be okay if you're going to listen to a psychologist do therapy with you, but fatal for a *self*-help group. Since the whole point of AA is being in a community of like-minded individuals, if you don't actually feel any personal connection to the project of quitting alcohol, it will just make you feel uncomfortable and out of place.

Also, uh, this just in, Brandsma didn't use a real AA group, because the real AA groups make people be anonymous which makes it inconvenient to research stuff. He just sort of started his own non-anonymous group, let's call it A, with no help from the rest of the fellowship, and had it do Alcoholics Anonymous-like stuff. On the other hand, many members of his control group went out into the community and…attended a real Alcoholics Anonymous, because Brandsma can't exactly ethically tell them not to. So *technically*, there were more people in AA in the no-AA group than in the AA group. Without knowing more about Alcoholics Anonymous, I can't know whether this objection is valid and whether Brandsma's group did or didn't capture the essence of the organization. Still, not the sort of thing you want to hear about a study.

Walsh et al (1991) is a similar study with similar confounders and similar results. Workers in an industrial plant who were in trouble for coming in drunk were randomly assigned either to an inpatient treatment program or to Alcoholics Anonymous. After a

year of followup, 60% of the inpatient-treated workers had stayed sober, but only 30% of the AA-treated workers had.

The pro-AA side made three objections to this study, of which one is bad and two are good.

The bad objection was that AA is cheaper than hospitalization, so even if hospitalization is good, AA might be more efficient – after all, we can't afford to hospitalize *everyone*. It's a bad objection because the authors of the study did the math and found out that hospitalization was so much better than AA that it decreased the level of further medical treatment needed and saved the health system more money than it cost.

The first good objection: like the Brandsma study, this study uses people under coercion – in this case, workers who would lose their job if they refused. Fine.

The second good objection, and this one is really interesting: *a lot of inpatient hospital rehab is AA*. That is, when you go to an hospital for inpatient drug treatment, you attend AA groups every day, and when you leave, they make you keep going to the AA groups. In fact, the study says that "at the 12 month and 24 month assessments, the rates of AA affiliation and attendance in the past 6 months did not differ significantly among the groups." Given that the hospital patients got hospital AA + regular AA, they were actually getting *more* AA than the AA group!

So all that this study proves is that AA + more AA + other things is better than AA. There was no "no AA" group, which makes it impossible to discuss how well AA does or doesn't work. Frick.

Timko (2006) is the only study I can hesitantly half-endorse. This one has a sort of clever methodological trick to get around the limitation that doctors can't ethically refuse to refer alcoholics to treatment. In this study, researchers at a Veterans' Affairs hospital randomly assigned alcoholic patients to "referral" or "intensive referral". In "referral", the staff asked the patients to go to AA. In "intensive referral", the researchers asked REALLY NICELY for the patients to go to AA, and gave them nice glossy brochures on how great AA was, and wouldn't shut up about it, and arranged for them to meet people at their first AA meeting so they could have friends in AA, et cetera, et cetera. The hope was that more people in the "intensive referral" group would end out in AA, and ~~that indeed happened~~ scratch that, I just re-read the study and the same number of people in both groups went to AA and the intensive group actually completed a lower number of the 12 Steps on average, have I mentioned I hate all research and this entire field is terrible? But the intensive referral people were more likely to have "had a spiritual awakening" and "have a sponsor", so it was decided the study wasn't a complete loss and when it was found the intensive referral condition had slightly less alcohol use the authors decided to declare victory.

So, whereas before we found that AA + More AA was better than AA, and that proved AA didn't work, in this study we find that AA + More AA was better than AA, and that proves AA *does* work. You know, did I say I hesitantly half-endorsed this study? Scratch that. I hate this study too.

IV.

All right, @#%^ this $@!&*. We need a *real* study, everything all lined up in a row, none of this garbage. Let's just hire half the substance abuse scientists in the country, throw a gigantic wad of money at them, give them as many patients as they need, let

them take as long as they want, but barricade the doors of their office and not let them out until they've proven something important beyond a shadow of a doubt.

This was about how the scientific community felt in 1989, when they launched Project MATCH. This eight-year, $30 million dollar, multi-thousand patient trial was supposed to solve everything.

The people going into Project MATCH might have been a little overconfident. Maybe "not even *Zeus* could prevent this study from determining the optimal treatment for alcohol addiction" overconfident. This might have been a mistake.

The study was designed with three arms, one for each of the popular alcoholism treatments of the day. The first arm would be "twelve step facilitation", a form of therapy based off of Alcoholics Anonymous. The second arm would be cognitive behavioral therapy, the most bog-standard psychotherapy in the world and one which by ancient tradition must be included in any kind of study like this. The third arm would be motivational enhancement therapy, which is a very short intervention where your doctor tells you all the reasons you should quit alcohol and tries to get you to convince yourself.

There wasn't a "no treatment" arm. This is where the overconfidence might have come in. Everyone knew alcohol treatment *worked*. Surely you couldn't dispute *that*. They just wanted to see which treatment worked best for which people. So you would enroll a bunch of different people – rich, poor, black, white, married, single, chronic alcoholic, new alcoholic, highly motivated, unmotivated – and see which of these people did best in which therapy. The result would be an algorithm for deciding where to send each of your patients. Rich black single chronic unmotivated alcoholic? We've found with $p < 0.00001$ that the best place for someone like that is in motivational enhancement therapy. Such was the dream. So, eight years and thirty million dollars and the careers of several prestigious researchers later, the results come in, and - yeah, everyone does exactly the same on every kind of therapy (with one minor, possibly coincidental exception). Awkward. "Everybody has won and all must have prizes!". If you're an optimist, you can say all treatments work and everyone can keep doing whatever they like best. If you're a pessimist, you might start wondering whether anything works at all.

By my understanding this is also the confusing conclusion of Ferri, Amato & Davoli (2006), the Cochrane Collaboration's attempt to get in on the AA action. Like all Cochrane Collaboration studies since the beginning of time, they find there is insufficient evidence to demonstrate the effectiveness of the intervention being investigated. This has been oft-quoted in the anti-AA literature. But by my reading, they had no control groups and were comparing AA to different types of treatment:

> Three studies compared AA combined with other interventions against other treatments and found few differences in the amount of drinks and percentage of drinking days. Severity of addiction and drinking consequence did not seem to be differentially influenced by TSF versus comparison treatment interventions, and no conclusive differences in treatment drop out rates were reported.

So the two best sources we have – Project MATCH and Cochrane – don't find any significant differences between AA and other types of therapy. Now, to be fair, the inpatient treatment mentioned in Walsh et al wasn't included, and inpatient treatment might be the gold standard here. But sticking to various forms of outpatient intervention, they all seem to be about the same.

So, the $64,000 question: do all of them work well, or do all of them work poorly?

V.

Alcoholism studies avoid control groups like they are on fire, presumably because it's unethical not to give alcoholics treatment or something. However, there is one class of studies that doesn't have that problem. These are the ones on "brief opportunistic intervention", which is much like a turbocharged even shorter version of "motivational enhancement therapy". Your doctor tells you 'HELLO HAVE YOU CONSIDERED QUITTING ALCOHOL??!!' and sees what happens.

Brief opportunistic intervention is the most trollish medical intervention ever, because here are all these brilliant psychologists and counselors trying to unravel the deepest mysteries of the human psyche in order to convince people to stop drinking, and then someone comes along and asks "Hey, have you tried just asking them politely?". And it works.

Not consistently. But it works for about one in eight people. And the theory is that since it only takes a minute or two of a doctor's time, it scales a lot faster than some sort of hideously complex hospital-based program that takes thousands of dollars and dozens of hours from everyone involved. If doctors would just spend five minutes with each alcoholic patient reminding them that no, really, alcoholism is really bad, we could cut the alcoholism rate by 1/8.

(this also works for smoking, by the way. I do this with every single one of my outpatients who smoke, and most of the time they roll their eyes, because their doctor is giving them *that speech*, but every so often one of them tells me that yeah, I'm right, they know they really should quit smoking and they'll give it another try. I have never saved anyone's life by dramatically removing their appendix at the last possible moment, but I have gotten enough patients to promise me they'll try quitting smoking that I think I've saved at least one life just by obsessively doing brief interventions every chance I get. This is probably *the* most effective life-saving thing you can do as a doctor, enough so that if you understand it you *may* be licensed to ignore 80,000 Hours' arguments on doctor replaceability)

Anyway, for some reason, it's okay to do these studies with control groups. And they are so fast and easy to study that everyone studies them all the time. A meta-analysis of 19 studies is unequivocal that they definitely work.

Why do these work? My guess is that they do two things. First, they hit people who honestly didn't realize they had a problem, and inform them that they do. Second, the doctor usually says they'll "follow up on how they're doing" the next appointment. This means that a respected authority figure is suddenly monitoring their drinking and will glare at them if they stay they're still alcoholic. As someone who has gone into a panic because he has a dentist's appointment in a week and he hasn't been flossing enough – and then flossed until his teeth were bloody so the dentist wouldn't be disappointed – I can sympathize with this.

But for our purposes, the brief opportunistic intervention sets a lower bound. It says "Here's a really minimal thing that seems to work. Do other things work better than this?"

The "brief treatment" is the next step up from brief intervention. It's an hour-or-so-long session (or sometimes a couple such sessions) with a doctor or counselor where they tell you some tips for staying off alcohol. I bring it up here because the brief

treatment research community spends its time doing studies that show that brief treatments are just as good as much more intense treatments. This might be most comparable to the “motivational enhancement therapy” in the MATCH study.

Chapman and Huygens (1988) find that a single interview with a health professional is just as good as six weeks of inpatient treatment (I don’t know about their hospital in New Zealand, but for reference six weeks of inpatient treatment in *my* hospital costs about $40,000.)

Edwards (1977) finds that in a trial comparing “conventional inpatient or outpatient treatment complete with the full panoply of services available at a leading psychiatric institution and lasting several months” versus an hour with a doc, both groups do the same at one and two year followup.

And so on.

All of this is starting to make my head hurt, but it’s a familiar sort of hurt. It’s the way my head hurts when Scott Aaronson talks about complexity classes

. We have all of these different categories of things, and some of them are the same as others and others are bigger than others but we’re not sure exactly where all of them stand.

We have classes “no treatment”, “brief opportunistic intervention”, “brief treatment”, “Alcoholics Anonymous”, “psychotherapy”, and “inpatient”.

We can prove that BOI > NT, and that AA = PT. Also that BT = IP = PT. We also have that IP > AA, which unfortunately we can use to prove a contradiction, so let’s throw it out for now.

So the hierarchy of classes seems to be (NT) < (BOI) ? (BT, IP, AA, PT) - in other words, no treatment is the worst, brief opportunistic intervention is better, and then *somewhere* in there we have this class of everything else that is the same.

Can we prove that BOI = BT?

We have some good evidence for this, once again from our *Handbook*. A study in Edinburgh finds that five minutes of psychiatrist advice (brief opportunistic intervention) does the same as sixty minutes of advice plus motivational interviewing (brief treatment).

So if we take all this seriously, then it looks like every psychosocial treatment (including brief opportunistic intervention) is the same, and all are better than no treatment. This is a common finding in psychiatry and psychology – for example, all common antidepressants are better than no treatment but work about equally well; all psychotherapies are better than no treatment but work about equally well, et cetera. It’s still an open question what this says about our science and our medicine.

The strongest counterexample to this is Walsh et al which finds the inpatient hospital stay works better than the AA referral, but this study looks kind of lonely compared to the evidence on the other side. And even the authors admit they were surprised by the effectiveness of the hospital there.

And let’s go back to Project MATCH. There wasn’t a control group. But there were the people who dropped out of the study, who said they’d go to AA or psychotherapy but never got around to it. Cutter and Fishbain (2005) take a look at what happened to these folks. They find that the dropouts did 75% as well as the people in any of the therapy groups, and that most of the effect of the therapy groups occurred in the first week (ie people dropped out after one week did about 95% as well as people who stayed in).

To me this suggests two things. First, therapy is only a little helpful over most people quitting on their own. Second, insofar as therapy is helpful, the tiniest brush with therapy is enough to make someone think “Okay, I’ve had some therapy, I’ll be better now”. Just like with the brief opportunistic interventions, five minutes of almost anything is enough.

This is a weird conclusion, but I think it’s the one supported by the data.

VI.

I should include a brief word about this giant table.

I see it everywhere. It looks very authoritative and impressive and, of course, giant. I believe the source is Miller’s Handbook of Alcoholism Treatment Approaches: Effective Alternatives, 3rd Edition

, the author of which is known as a very careful scholar whom I cannot help but respect.

And the table does a good thing in discussing medications like acamprosate and naltrexone, which are very important and effective interventions but which will not otherwise be showing up in this post.

However, the therapy part of the table looks really wrong to me.

First of all, I notice acupuncture is ranked 17 out of 48, putting in a much, *much* better showing than treatments like psychotherapy, counseling, or education. Seems fishy.

Second of all, I notice that motivational enhancement (#2), cognitive therapy (#13), and twelve-step (#37) are all about as far apart as could be, but the largest and most powerful trial ever, Project MATCH, found all three to be about equal in effectiveness.

Third of all, I notice that cognitive therapy is at #13, but psychotherapy is at #46. But cognitive therapy is a kind of psychotherapy.

Fourth of all, I notice that brief interventions, motivational enhancement, confrontational counseling, psychotherapy, general alcoholism counseling, and education are all over. But a lot of these are hard to differentiate from one another.

The table seems messed up to me. Part of it is because it is about evidence base rather than effectiveness (consider that handguns have a stronger evidence base than the atomic bomb, since they have been used many more times in much better controlled conditions, but the atomic bomb is more effective) and therefore acupuncture, which is poorly studied, can rank quite high compared to things which have even one negative study.

But part of it just seems wrong. I haven't read the full book, but I blame the tendency to conflate studies showing "X does not work better than anything else" with "X does not work".

Remember, whenever there are meta-analyses that contradict single very large well-run studies, go with the single very large well-run study, especially when the meta-analysis is as weird as this one. Project MATCH is the single very large well-run study, and it says this is balderdash. I'm guessing it's trying to use some weird algorithmic methodology to automatically rate and judge each study, but that's no substitute for careful human review.

VII.

In conclusion, as best I can tell – and it is not very well, because the studies that could really prove anything robustly haven't been done – most alcoholics get better on their own. All treatments for alcoholism, including Alcoholics Anonymous, psychotherapy, and just a few minutes with a doctor explaining why she thinks you need to quit, increase this already-high chance of recovery a small but nonzero amount. Furthermore, they are equally effective after only a tiny dose: your first couple of meetings, your first therapy session. Some studies suggest that inpatient treatment with outpatient followup may be better than outpatient treatment alone, but other studies contradict this and I am not confident in the assumption.

So does Alcoholics Anonymous work? Though I cannot say anything authoritatively, my impression is: Yes, but only a tiny bit, and for many people five minutes with a doctor may work just as well as years completing the twelve steps. As such, individual alcoholics may want to consider attending if they don't have easier options; doctors might be better off just talking to their patients themselves.

If this is true – and right now I don't have much confidence that it is, it's just a direction that weak and contradictory data are pointing – it would be really awkward for the multibazillion-dollar treatment industry.

More worrying, I am afraid of what it would do to the War On Drugs. Right now one of the rallying cries for the anti-Drug-War movement is "treatment, not prison". And although I haven't looked seriously at the data for any drug besides alcohol. I think some data there are similar. There's very good medication for drugs – for example methadone and suboxone for opiate abuse – but in terms of psychotherapy it's mostly the same stuff you get for alcohol. Rehabs, whether they work or not, seem to serve an important sort of ritual function, where if you can send a drug abuser to a rehab you at least feel like something has been done. Deny people that ritual, and it might make prison the only politically acceptable option.

In terms of things to actually treat alcoholism, I remain enamoured of the Sinclair Method, which has done crazy outrageous stuff like conduct an experiment *with an actual control group*. But I haven't investigated enough to know whether my early excitement about them looks likely to pan out or not.

I would not recommend quitting any form of alcohol treatment that works for you, or refusing to try a form of treatment your doctor recommends, based on any of this information.

Prescriptions, Paradoxes, and Perversities

[WARNING: I am not a pharmacologist. I am not a researcher. I am not a statistician. This is not medical advice. This is really weird and you should not take it too seriously until it has been confirmed]

I.

I've been playing around with data from Internet databases that aggregate patient reviews of medications.

Are these any good? I looked at four of the largest such databases – Drugs.com, WebMD, AskAPatient, and DrugLib – as well as psychiatry-specific site CrazyMeds – and took their data on twenty-three major antidepressants. Then I correlated them with one another to see if the five sites mostly agreed.

Correlations between Drugs.com, AskAPatient, and WebMD were generally large and positive (around 0.7). Correlations between CrazyMeds and DrugLib were generally small or negative. In retrospect this makes sense, because these two sites didn't allow separation of ratings by condition, so for example Seroquel-for-depression was being mixed with Seroquel-for-schizophrenia.

So I threw out the two offending sites and kept Drugs.com, AskAPatient, and WebMD. I normalized all the data, then took the weighted average of all three sites. From this huge sample (the least-reviewed drug had 35 ratings, the most-reviewed drug 4,797) I obtained a unified opinion of patients' favorite and least favorite antidepressants.

This doesn't surprise me at all. Everyone secretly knows Nardil and Parnate (the two commonly-used drugs in the MAOI class) are excellent antidepressants[1]. Oh, nobody will prescribe them, because of the dynamic discussed here, but in their hearts they know it's true.

Likewise, I feel pretty good to see that Serzone, which I recently defended, is number five. I've had terrible luck with Viibryd, and it just seems to make people taking it more annoying, which is not a listed side effect but which I swear has happened.

The table also matches the evidence from chemistry – drugs with similar molecular structure get similar ratings, as do drugs with similar function. This is, I think, a good list.

Which is too bad, because it makes the next part that much more terrifying.

II.

There is a sixth major Internet database of drug ratings. It is called RateRx, and it differs from the other five in an important way: it solicits ratings from doctors, not patients. It's a great idea – if you trust your doctor to tell you which drug is best, why not take advantage of wisdom-of-crowds and trust *all* the doctors?

The RateRX logo. Spoiler: this is going to seem really ironic in about thirty seconds.

RateRx has a modest but respectable sample size – the drugs on my list got between 32 and 70 doctor reviews. There's only one problem.

You remember patient reviews on the big three sites correlated about +0.7 with each other, right? So patients pretty much agree on which drugs are good and which are bad?

Doctor reviews on RateRx correlated at -0.21 with patient reviews. The negative relationship is nonsignificant, but that just means that at best, doctor reviews are totally uncorrelated with patient consensus.

This has an obvious but very disturbing corollary. I couldn't get good numbers on how times each of the antidepressants on my list were prescribed, because the information I've seen only gives prescription numbers for a few top-selling drugs, plus we've got the same problem of not being able to distinguish depression prescriptions from anxiety prescriptions from psychosis prescriptions. But total number of online reviews makes a pretty good proxy. After all, the more patients are using a drug, the more are likely to review it.

Quick sanity check: the most reviewed drug on my list was Cymbalta. Cymbalta was also the best selling antidepressant of 2014. Although my list doesn't exactly track the best-sellers, that seems to be a function of how long a drug has been out – a best-seller that came out last year might have only 1/10th the number of reviews as a best-seller that came out ten years ago. So number of reviews seems to be a decent correlate for amount a drug is used.

In that case, amount a drug is used correlates highly (+0.67, p = 0.005) with doctors' opinion of the drug, which makes perfect sense since doctors are the ones prescribing it. But amount the drug gets used correlates negatively with patient rating of the drug (-0.34, p = ns), which of course is to be expected given the negative correlation between doctor opinion and patient opinion.

So the more patients like a drug, the less likely it is to be prescribed[2].

III.

There's one more act in this horror show.

Anyone familiar with these medications reading the table above has probably already noticed this one, but I figured I might as well make it official.

I correlated the average rating of each drug with the year it came on the market. The correlation was -0.71 (p < .001). That is, the newer a drug was, the less patients liked it[3].

This pattern absolutely *jumps* out of the data. First- and second- place winners Nardil and Parnate came out in 1960 and 1961, respectively; I can't find the exact year third-place winner Anafranil came out, but the first reference to its trade name I can find in the literature is from 1967, so I used that. In contrast, last-place winner Viibryd came

out in 2011, second-to-last place winner Abilify got its depression indication in 2007, and third-to-last place winner Brintellix is as recent as 2013.

This result is robust to various different methods of analysis, including declaring MAOIs to be an unfair advantage for Team Old and removing all of them, changing which minor tricylics I do and don't include in the data, and altering whether Deprenyl, a drug that technically came out in 1970 but received a gritty reboot under the name Emsam in 2006, is counted as older or newer.

So if you want to know what medication will make you happiest, at least according to this analysis your best bet isn't to ask your doctor, check what's most popular, or even check any individual online rating database. It's to look at the approval date on the label and choose the one that came out first.

IV.

What the *hell* is going on with these data?

I would like to dismiss this as confounded, but I have to admit that any reasonable person would expect the confounders to go the opposite way.

That is: older, less popular drugs are usually brought out only when newer, more popular drugs have failed. MAOIs, the clear winner of this analysis, are very clearly reserved in the guidelines for "treatment-resistant depression", ie depression you've already thrown everything you've got at. But these are precisely the depressions that are hardest to treat.

Imagine you are testing the fighting ability of three people via ten boxing matches. You ask Alice to fight a Chihuahua, Bob to fight a Doberman, and Carol to fight Cthulhu. You would expect this test to be biased in favor of Alice and against Carol. But MAOIs and all these other older rarer drugs are practically never brought out except against Cthulhu. Yet they *still* have the best win-loss record.

Here are the only things I can think of that might be confounding these results.

Perhaps because these drugs are so rare and unpopular, psychiatrists only use them when they have really really good reason. That is, the most popular drug of the year they pretty much cluster-bomb everybody with. But every so often, they see some patient who seems absolutely 100% perfect for clomipramine, a patient who practically *screams* "clomipramine!" at them, and then they give this patient clomipramine, and she does really well on it.

(but psychiatrists aren't actually that good at personalizing antidepressant treatments. The only thing even *sort of* like that is that MAOIs are extra-good for a subtype called atypical depression. But that's like a third of the depressed population, which doesn't leave much room for this super-precise-targeting hypothesis.)

Or perhaps once drugs have been on the market longer, patients figure out what they like. Brintellix is so new that the Brintellix patients are the ones whose doctors said "Hey, let's try you on Brintellix" and they said "Whatever". MAOIs have been on the market so long that presumably MAOI patients are ones who tried a dozen antidepressants before and stayed on MAOIs because they were the only ones that worked.

(but Prozac has been on the market 25 years now. This should only apply to a couple of very new drugs, not the whole list.)

Or perhaps the older drugs have so many side effects that no one would stay on them unless they're absolutely perfect, whereas people are happy to stay on the newer drugs even if they're not doing much because whatever, it's not like they're causing any trouble.

(but Seroquel and Abilify, two very new drugs, have awful side effects, yet are down at the bottom along with all the other new drugs)

Or perhaps patients on very rare weird drugs get a special placebo effect, because they feel that their psychiatrist cares enough about them to personalize treatment. Perhaps they identify with the drug – "I am special, I'm one of the only people in the world who's on nefazodone!" and they become attached to it and want to preach its greatness to the world.

(but drugs that are rare because they are especially new don't get that benefit. I would expect people to also get excited about being given the latest, flashiest thing. But only drugs that are rare because they are old get the benefit, not drugs that are rare because they are new.)

Or perhaps psychiatrists tend to prescribe the drugs they "imprinted on" in medical school and residency, so older psychiatrists prescribe older drugs and the newest psychiatrists prescribe the newest drugs. But older psychiatrists are probably much more experienced and better at what they do, which could affect patients in other ways – the placebo effect of being with a doctor who radiates competence, or maybe the more experienced psychiatrists are really good at psychotherapy, and that makes the patient better, and they attribute it to the drug.

(but read on…)

V.

Or perhaps we should take this data at face value and assume our antidepressants have been getting worse and worse over the past fifty years.

This is not entirely as outlandish as it sounds. The history of the past fifty years has been a history of moving from drugs with more side effects to drugs with fewer side effects, with what I consider somewhat less than due diligence in making sure the drugs were quite as effective in the applicable population. This is a *very* complicated and controversial statement which I will be happy to defend in the comments if someone asks.

The big problem is: drugs go off-patent after twenty years. Drug companies want to push new, on-patent medications, and most research is funded by drug companies. So lots and lots of research is aimed at proving that newer medications invented in the past twenty years (which make drug companies money) are better than older medications (which don't).

I'll give one example. There is only a single study in the entire literature directly comparing the MAOIs – the very old antidepressants that did best on the patient ratings – to SSRIs, the antidepressants of the modern day[4]. This study found that phenelzine, a typical MAOI, was no better than Prozac, a typical SSRI. Since Prozac had fewer side effects, that made the choice in favor of Prozac easy.

Did you know you can look up the authors of scientific studies on LinkedIn and sometimes get very relevant information? For example, the lead author of this study has a resume that clearly lists him as working for Eli Lilly at the time the study was conducted (spoiler: Eli Lilly is the company that makes Prozac). The second author's LinkedIn profile shows he is *also* an operations manager for Eli Lilly. Googling the fifth author's name links to a news article about Eli Lilly making a $750,000 donation to his clinic. Also there's a little blurb at the bottom of the paper saying "Supported by a research grant by Eli Lilly and company", then thanking several Eli Lilly executives by name for their assistance.

This is the sort of study which I kind of wish had gotten replicated *before* we decided to throw away an entire generation of antidepressants based on the result.

But who will come to phenelzine's defense? Not Parke-Davis , the company that made it: their patent expired sometime in the seventies, and then they were bought out by Pfizer[5]. And not Pfizer – without a patent they can't make any money off Nardil, and besides, Nardil is competing with their own on-patent SSRI drug Zoloft, so Pfizer has as much incentive as everyone else to push the "SSRIs are best, better than all the rest" line.

Every twenty years, pharmaceutical companies have an incentive to suddenly declare that all their old antidepressants were awful and you should never use them, but whatever new antidepressant they managed to dredge up is super awesome and you should use it all the time. This sort of *does* seem like the sort of situation that might lead to older medications being better than newer ones. A couple of people have been pushing this line for years – I was introduced to it by Dr. Ken Gillman from Psychotropical Research, whose recommendation of MAOIs and Anafranil as most effective match the patient data very well, and whose essay Why Most New Antidepressants Are Ineffective is worth a read.

I'm not sure I go as far as he does – even if new antidepressants aren't worse outright, they might still trade less efficacy for better safety. Even if they handled the tradeoff well, it would look like a net loss on patient rating data. After all, assume Drug A is 10% more effective than Drug B, but also kills 1% of its users per year, while Drug B kills nobody. Here there's a good case that Drug B is much better and a true advance. But Drug A's ratings would look better, since dead men tell no tales and don't get to put their objections into online drug rating sites. Even if victims' families did give the drug the lowest possible rating, 1% of people giving a very low rating might still not counteract 99% of people giving it a higher rating.

And once again, I'm not sure the tradeoff is handled very well at all.[6].

VI.

In order to distinguish between all these hypotheses, I decided to get a lot more data.

I grabbed all the popular antipsychotics, antihypertensives, antidiabetics, and anticonvulsants from the three databases, for a total of 55,498 ratings of 74 different drugs. I ran the same analysis on the whole set.

The three databases still correlate with each other at respectable levels of +0.46, +0.54, and +0.53. All of these correlations are highly significant, $p < 0.01$. The negative correlation between patient rating and doctor rating remains and is now a

highly significant -0.344, $p < 0.01$. This is robust even if antidepressants are removed from the analysis, and is notable in both psychiatric and nonpsychiatric drugs.

The correlation between patient rating and year of release is a no-longer-significant -0.191. This is heterogenous; antidepressants and antipsychotics show a strong bias in favor of older medications, and antidiabetics, antihypertensives, and anticonvulsants show a slight nonsignificant bias in favor of newer medications. So it would seem like the older-is-better effect is purely psychiatric.

I conclude that for some reason, there really is a highly significant effect across all classes of drugs that makes doctors love the drugs patients hate, and vice versa.

I also conclude that older psychiatric drugs seem to be liked much better by patients, and that this is not some kind of simple artifact or bias, since if such an artifact or bias existed we would expect it to repeat in other kinds of drugs, which it doesn't.

VII.

Please feel free to check my results. Here is a spreadsheet (.xls) containing all of the data I used for this analysis. Drugs are marked by class: 1 is antidepressants, 2 is antidiabetics, 3 is antipsychotics, 4 is antihypertensives, and 5 is anticonvulsants. You should be able to navigate the rest of it pretty easily.

One analysis that needs doing is to separate out drug effectiveness versus side effects. The numbers I used were combined satisfaction ratings, but a few databases – most notably WebMD – give you both separately. Looking more closely at those numbers might help confirm or disconfirm some of the theories above.

If anyone with the necessary credentials is interested in doing the hard work to publish this as a scientific paper, drop me an email and we can talk.

Footnotes

1. Technically, MAOI superiority has only been proven for atypical depression, the type of depression where you can still have changing moods but you are unhappy on net. But I'd speculate that right now most patients diagnosed with depression have atypical depression, far more than the studies would indicate, simply because we're diagnosing less and less severe cases these days, and less severe cases seem more atypical.

2. First-place winner Nardil has only 16% as many reviews as last-place winner Viibryd, even though Nardil has been on the market fifty years and Viibryd for four. Despite its observed superiority, Nardil may very possibly be prescribed less than 1% as often as Viibryd.

3. Pretty much the same thing is true if, instead of looking at the year they came out, you just rank them in order from earliest to latest.

4. On the other hand, what we do have is a lot of studies comparing MAOIs to imipramine, and a lot of other studies comparing modern antidepressants to imipramine. For atypical depression and dysthymia, MAOIs beat imipramine handily, but the modern antidepressants are about equal to imipramine. This strongly implies the MAOIs beat the modern antidepressants in these categories.

5. Interesting Parke-Davis facts: Parke-Davis got rich by being the people to market cocaine back in the old days when people treated it as a pharmaceutical, which must have been kind of like a license to print money. They also worked on hallucinogens with no less a figure than Aleister Crowley, who got a nice tour of their facilities in Detroit.

6. Consider: *Seminars In General Psychiatry* estimates that MAOIs kill one person per 100,000 patient years. A third of all depressions are atypical. MAOIs are 25 percentage points more likely to treat atypical depression than other antidepressants. So for every 100,000 patients you give a MAOI instead of a normal antidepressant, you kill one and cure 8,250 who wouldn't otherwise be cured. The QALY database says that a year of moderate depression is worth about 0.6 QALYs. So for every 100,000 patients you give MAOIs, you're losing about 30 QALYs and gaining about 3,300.

Guns And States

[Epistemic status: I think I probably wrung the right conclusions out of this evidence, but this isn't the only line of evidence bearing on the broader gun control issue and all I can say is what it's consistent with. Content warning for discussion of suicide, murder, and race]

I.

From a Vox article on America's Gun Problem, Explained: "On Wednesday, it happened again: There was a mass shooting — this time, in San Bernardino, California. And once again on Sunday, President Barack Obama called for measures that make it harder for would-be shooters to buy deadly firearms."

Then it goes on to say that "more guns mean more gun deaths, period. The research on this is overwhelmingly clear. No matter how you look at the data, more guns mean more gun deaths." It cites the following chart:

...then uses the graph as a lead in to talk about active shooter situations, gun-homicide relationships, and outrage over gun massacres.

Did you notice that the axis of this graph says "gun deaths", and that this is a totally different thing from gun murders?

(this isn't an isolated incident: Vox does the same thing here and here)

Gun deaths are a combined measure of gun homicides and gun suicides. Here is a graph of guns vs. gun homicides:

And here is a graph of guns vs. gun suicides:

The relationship between gun ownership and homicide is weak (and appears negative), the relationship between gun ownership and suicide is strong and positive. The entire effect Vox highlights in their graph is due to gun suicides, but they are using it to imply conclusions about gun homicides. This is why you shouldn't make a category combining two unlike things.

II.

I am not the first person to notice this. The Washington Examiner makes the same criticism of Vox's statistics that I do. And Robert VerBruggen of National Review does the same analysis decomposing gun deaths into suicides and homicides, and like me finds no correlation with homicides.

German Lopez of Vox responds here. He argues that VerBruggen can't just do a raw uncontrolled correlation of state gun ownership with state murder rates without adjusting for confounders. This is true, although given that Vox has done this time and time again for months on end and all VerBruggen is doing is correctly pointing out a flaw in their methods, it feels kind of like an isolated demand for rigor.

So let's look at the more-carefully-controlled studies. Lopez suggests the ones at the Harvard Injury Control Research Center, which has done several statistical analyses of gun violence. They list two such analyses comparing gun ownership versus homicide rates across US states: Miller Azrael & Hemenway (2002), and Miller Azrael & Hemenway (2007).

(does it count as nominative determinism when someone named Azrael goes into homicide research?)

We start with MA&H 2002. This study does indeed conclude that higher gun ownership rates are correlated with higher murder rates after adjusting for confounders. But suspiciously, it in fact finds that higher gun ownership rates are correlated with higher murder rates even *before* adjusting for confounders, something that we already found wasn't true! Furthermore, even after adjusting for confounders it finds in several age categories that higher gun ownership rates are correlated with higher *non-gun homicide rates* (eg the rates at which people are murdered by knives or crowbars or whatever) at p less than 0.001. This is really suspicious! Unless guns are exerting some kind of malign pro-murder influence that makes people commit more knife murders, some sort of confounding influence has remained. Let's look closer.

The study gets its murder rate numbers from the National Center for Health Statistics, which seems like a trustworthy source. It gets its gun ownership numbers from…oh, that's interesting, it doesn't actually have any gun ownership numbers. It says that there is no way to figure out what percent of people in a given state own guns, so as a proxy for gun ownership numbers, it will use a measure called FS/S, ie the number of firearm suicides in a state divided by the total number of suicides.

This makes some intuitive sense. Among people who want to commit suicide, suppose a fixed percent prefer to use guns compared to other methods. In that case, the determining factor for whether or not they use a gun will be whether or not they have a gun. Hospitals diligently record statistics about suicide victims including method of suicide, so if our assumption holds this should be a decent proxy for gun ownership within a state.

There's only one problem – I checked this against an actual measure of gun ownership per state that came out after this study was published – the CDC asking 200,000 people how many guns they had as part of the Behavioral Risk Factor Surveillance System Survey – and the FS/S measure fails. When I repeat all of their analyses with their own FS/S measure, I get all of their same positive correlations, including the ones with non-gun homicides. When I repeat it with the real gun ownership data, all of these positive correlations disappear. When I look at exactly why this happens, it's because FS/S is *much* more biased towards Southern states than actual gun ownership is. Real gun ownership correlates very modestly – 0.25 – with 538's ranking of the Southern-ness of states. FS/S correlates at a fantastically high 0.62. For some reason, suicidal Southerners are much more likely to kill themselves with guns than suicidal people from the rest of the States, even when you control for whether they have a gun or not. That means that MA&H 2002 thought it was measuring gun ownership, but was actually measuring Southern-ness. This is why they found higher homicide rates, including higher rates of non-gun homicide.

So we move on to MA&H 2007. This study was published after the CDC's risk survey, so they have access to the same superior gun ownership numbers I used to pick apart their last study. They also have wised up to the fact that Southern-ness is important, and they include a dummy variable for it in their calculations. They also control for

non-gun crime rate, Gini coefficient, income, and alcohol use. They do *not* control for urbanization level or race, but when I re-analyze their data including these factors doesn't change anything, likely because they are already baked in to the crime rate.

They find that even after controlling for all of this stuff, there is still a significant correlation between gun ownership level and gun homicide rate. Further, this time they are using good statistics, and there is *not* a significant correlation between gun ownership and non-gun-homicide rate. Further, there *is* a correlation between gun ownership and total homicide rate, suggesting that the gun-gun-homicide correlation was not just an artifact of people switching from inferior weapons to guns while still committing the same number of murders. Further, this is robust to a lot of different decisions about what to control or not to control, and what to include or not to include.

I repeated all of their analyses using two different sources of gun ownership data, a couple different sources of homicide and crime rate data, and a bunch of different plausible and implausible confounders – thanks a lot to Tumblr user su3su2u1 for walking me through some of the harder analyses. I was able to replicate their results. Pro-gun researcher John Lott had many complaints about this study, including that it was insensitive to including DC and that it was based entirely on the questionable choice of controlling for robbery rate – but I was unable to replicate his concerns and found that the guns-homicide correlation remained even after DC was included and even when I chose a group of confounders not including robbery rate. I was unable to use their methodology to replicate the effect in places where it shouldn't replicate (I tried to convince it to tell me tractors caused homicide, since I was suspicious that it was just picking up an urban/rural thing, but it very appropriately refused to fall for it). Overall I am about as sure of this study as I have ever been of any social science study, ie somewhat.

This study doesn't prove causation; while one interpretation is that guns cause homicide, another is that homicide causes guns – for example, by making people feel unsafe so they buy guns to protect themselves. However, I doubt the reverse causation aspect in this case. The study controlled for robbery rate; ie it was looking at whether guns predicted homicides above and beyond those that could be expected given the level of non-homicide crime. My guess is that people feeling unsafe is based more on the general crime rate than on the homicide rate per se, which would make it hard for the homicide rate to cause increased gun ownership independently of the crime rate.

If guns are in fact correlated with more homicide, how come me and VerBruggen found the opposite in our simpler scatterplot analysis? This is complicated, but I think the biggest part of the answer is the urban/rural divide. Rural people have more guns. Murder rates are higher in urban areas. Race also plays a part: whites have more guns, but black areas have higher murder rates. Finally, the North and West seem to have more guns, but murder rates are highest in the South (which is what produced the bogus effect on the last study). All of these differences are large enough to cancel out the gun/no-gun difference and make the raw scatterplot look like nothing. This study didn't address all those things directly, but its decision to control for non-gun crime rate and poverty took care of them nevertheless. As the old saying goes, guns don't kill people; guns controlled for robbery rate, alcoholism, income, a dummy variable for Southernness, and a combined measure of social deprivation kill people.

If this is all true, how come I spent so much time yelling at that first study with worse data? Because I worry that if people only see the good studies, they'll get complacent. Vox posted these two studies as proof that there was a state-level gun-murder

correlation. The first one was deeply flawed, but the second one turned out to be okay. Do you think Vox realized this? Do you think they would have written that article any differently in a world where *both* studies were flawed? As long as you trust every scientific paper you see – let alone every scientific paper you see on your side in a highly politicized field – even when you're right it will often just be by luck.

III.

Vox also voxsplains to us about America's unusually high gun homicide rate.

Having presented this graph, they say that "To understand why that is, there's another important statistic: The US has by far the highest number of privately owned guns in the world."

Even granting, as we saw above, that gun ownership does indeed increase homicide rates, this is not the most important factor in explaining America's higher homicide rate, or even close to the most important factor. Let me give a few arguments for why this must be the case:

1. The United States' homicide rate of 3.8 is clearly higher than that of eg France (1.0), Germany (0.8), Australia (1.1), or Canada (1.4). However, as per the FBI, only 11,208 of our 16,121 murders were committed with firearms, eg 69%. By my calculations, that means our nonfirearm murder rate is 1.2. In other words, our *non-firearm homicide rate alone* is higher than France, Germany, and Australia's *total* homicide rate. Nor does this mean that if we banned all guns we would go down to 1.2 – there is likely a substitution effect where some murderers are intent on murdering and would prefer to use convenient firearms but will switch to other methods if they have to. 1.2 should be considered an absolute lower bound. And it is *still* higher than the countries we want to compare ourselves to.

2. There are many US states that combine very high firearm ownership with very low murder rates. The highest gun-ownership state in the nation is Wyoming, where 59.7% of households have a gun (really!). But Wyoming has a murder rate of only 1.4 – the same as right across the border in more gun-controlled Canada, and only about a third of that of the nation as a whole. It seems likely that the same factors giving Canada a low murder rate give Wyoming a low murder rate, and that the factors differentiating the rest of America from Wyoming are the same factors that differentiate the rest of America from Canada (and Germany, and France...). But this does not include lower gun ownership.

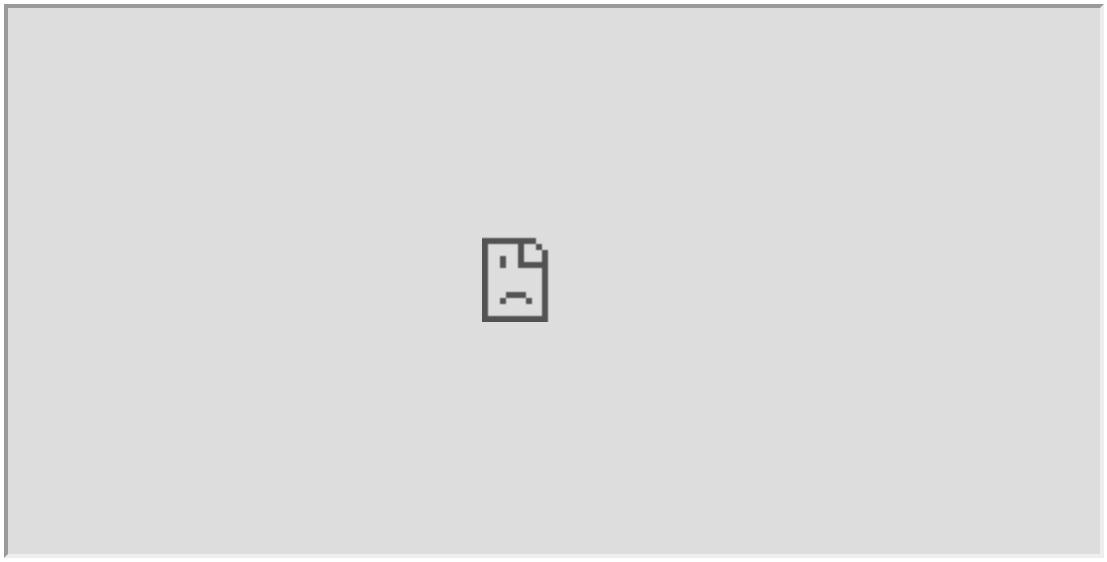
3. There are many US states that combine very low firearm ownership with very high murder rates. The highest murder rate in the country is that of Washington, DC, which has a murder rate of 21.8, more than twenty times that of most European countries. But DC also has the strictest gun bans and the *lowest* gun ownership rate in the country, with gun ownership numbers less than in many European states! It seems likely that the factors making DC so deadly are part of the story of why America as a whole is so deadly, but these cannot include high gun ownership.

If not gun ownership, what *is* the factor making America so much more deadly than Europe and other First World countries? The traditional answer I always heard to this question was that America had a "culture of violence". I always hated this answer, because it seemed so vague and meaningless as to be untestable by design. If the

NRA waves their hands and says “eh, culture of violence”, how are you going to tell them they’re wrong?

But we can work with this if we assume the culture of violence (or, if you want to be official about it, “honor culture”) is more common in some populations and areas than others. Some of the groups most frequently talked about during these lines are Southerners and various nonwhite minorities. This provides a testable theory: if we compare American non-Southern whites to European countries mostly made up of non-Southern whites, we’ll find similar murder rates. But first, some scatter plots:

This is murder rate by state, correlated with perceived Southernness of that state as per 538’s poll. I’ve removed DC as an outlier on all of the following.



And this is murder rate by state correlated with percent black population:

This would seem to support the “culture of violence” theory.

Can we adjust for this and see what the murder rate is for non-Southern whites? Sort of. The Economist gives a white-only murder rate of 2.5 (this is based on white victims, whereas we probably want white perpetrators, but the vast majority of murders are within-race so it doesn’t make much difference). And Audacious Epigone has put together a collection of white murder rates by state. I can’t find anything on non-Southern white murder rates per se, but one hack would be to take the white murder rate in non-Southern states and assume there aren’t any Southerners there.

Our main confounder will be urbanization. Western Europe is about 80% urban, so let’s look at states at a similar level. The four northern states that are closest to 80% urban are Colorado, Oregon, Washington, and Connecticut. I’m throwing out Colorado because it has a large Latino population who can’t be statistically differentiated from whites. That leaves, Washington (2.4), Connecticut (2.0), and Oregon (2.0). So possibly adjusting out Southerners brings us down from 2.5 (all whites) to 2.1 or so

(non-Southern whites)? Again, compare to Germany at 0.8, Canada at 1.4, and America at 3.8.

There's one more factor that needs to be considered:

This is a plot of the gun death rate vs. the robbery rate. There's a strong correlation (r = 0.78). Robbery is heavily correlated with percent black, percent Southern, and urbanization, so it's probably coming from the same place. Nevertheless, it seems to correlate with murder better than any of them alone, maybe because it's combining all three measures together. I was able to make a linear model using those three measures that correlated at r = 0.79 with murder, about the same amount that robbery does. I should also mention that robbery correlates *negatively* with gun ownership at r = – 0.52, but this disappeared when controlled for urbanization.

So my very tentative conclusion is that although the US murder rate is much higher than that of other First World countries, this is partly due to the existence of various cultural factors not present in those other nations. When we adjust those away, America's murder rate falls from 3.8 to 2.1. Which is still higher than Germany's 0.8 or Canada's 1.4.

Is that extra due to guns?

IV.

According to MA&H 2007, each *absolute* percentage point in gun ownership was related to a 2.2 *relative* percentage point difference in homicide. This part of the study was beyond my ability to check, and I'm not sure why they switched from absolute to relative percents there, but suppose we take it seriously.

America has a gun ownership rate of 32%, so if we somehow decreased that to zero, we would naively expect about a 70% decrease in homicides. Unfortunately, only 67% of American homicides involve guns, so we're back to pretending that eliminating guns will not only have zero substitution effect but also magically prevent non-gun homicides. This shows the dangers of extrapolating a figure determined by small local differences all the way to the edge of the graph (I'M TALKING TO YOU, RAY KURZWEIL).

Maybe we can be more modest? Canada has a gun ownership rate of aboot 26%, so…

…wait a second. I thought we've been told that the US has a gun ownership rate seven zillion times that of any other country in the world, and that is why we are so completely unique in our level of gun crime? And now they're telling us that Canada has 26% compared to our 32%? What?

Don't trust me too much here, because I've never seen anyone else analyze this and it seems like the sort of thing there should be loads of analyses of if it's true, but I think the difference is between percent of households with guns vs. guns per capita. US and Canada don't differ very much in percent of households with guns, but America has about four times as many guns per capita. Why? I have no idea, but the obvious implication is that Canadians mostly stop at one gun, whereas Americans with guns buy lots and lots of them. In retrospect this makes sense; I am looking at gun enthusiast bulletin boards, and they're advising other gun enthusiasts that six guns is really the bare minimum it's possible to get by with (see also "How many guns can you have before it's okay to call your collection an 'arsenal'?", which I have to admit is

not a question that I as a boring coastal liberal have ever considered). So if the guy asking that question decides he needs 100 guns before he gets his arsenal merit badge, that's a lot more guns per capita without increasing percent household gun ownership. This should actually be another argument that guns are not a major factor in differentiating US vs. Canadian murder rates, since unless you're going on a mass shooting (WHICH IS REALLY RARE) you wouldn't expect more murders from any gun in a household beyond the first. That means that the small difference between US and Canadian household percent gun ownership rates (32% vs. 26%) would have to drive the large difference between US and Canadian murder rates (1.4 vs. 3.8), which just isn't believable.

…okay, sorry, where were we? Canada has a gun ownership rate of about 26%, so if America were to get its gun ownership as low as Canada, that would be -6 absolute percentage points = a 13% relative decrease in murder rate = the murder rate going from 3.8 to 3.3 = a 0.5 point decrease in the murder rate. That's pretty close to the difference between our 2.1 US-sans-culture-of-violence estimate and the 1.4 Canadian rate – so maybe beyond the cultures of violence, the rest of the US/Canada difference really is due to guns?

(I'm not sure whether I should be subtracting 13% from 2.1 rather than 3.8 here)

In Germany, 9% of households own firearms (wait, really? European gun control is less strict than I thought!) Using MA&H's equation, we predict that if the US had the same gun ownership rate as Germany, its murder rate would drop 50%, eg from 3.8 to 1.9. Adjust out the culture of violence, and we're actually pretty close to real Germany's murder rate of 0.8.

How much would gun control actually cut US gun ownership? That obviously depends on the gun control, but a lot of people talk about Australia's gun buyback program as a model to be emulated. These people say it decreased gun ownership from 7% of people to 5% of people (why is this number so much lower than Canada and Germany? I think because it's people rather than households – if a gun owner is married to a non-gun-owner, they count as one gun-owner and one non-owner, as opposed to a single gun-owning household. The Australian household number seems to be 19% or so). So the gun buyback program in Australia decreased gun ownership by (relative) 30% or so. If a similar program decreased gun ownership in America by (relative) 30%, it would decrease it by (absolute) 10% and decrease the homicide rate by (absolute) 22%. Since there are about 13000 homicides in the US per year, that would save about 3000 lives – or avert about one 9/11 worth of deaths per year.

(note that our murder rate would still be 3.0, compared to Germany's 0.8 and Canada's 1.4. Seriously, I'm telling you, the murder rate difference is not primarily driven by guns!)

Is that worth it? That obviously depends on how much you like being able to have guns. But let me try to put this number into perspective in a couple of different ways:

Last time anyone checked, which was 1995, about 618,000 people died young (ie before age 65) in the US per year. Suppose that the vast majority of homicides are of people below 65. That means that instituting gun control would decrease the number of premature deaths to about 615,000 – in other words, by about half a percentage point. I'm having to borrow this data from the UK, but if it carries over, the average person my age (early 30s) has a 1/1850 chance of death each year. Gun control would

decrease that to about 1/1860. I'm very very unsure about the exact numbers, but it seems like the magnitude is very low.

On the other hand, lives are very valuable. In fact, the statistical value of a human life in the First World – ie the value that groups use to decide whether various life-saving interventions are worth it or not – is $7.4 million. That means that gun control would "save" $22 billion dollars a year. Americans buy about 20 million guns per year (really)! If we were to tax guns to cover the "externality" of gun homicides preventable by Australia-level gun control, we would have to slap a $1000 tax on each gun sold. While I have no doubt that some people, probably including our arsenal collector above, would be willing to pay that, my guess is that most people would not. This suggests that most people probably do not enjoy guns enough to justify keeping them around despite their costs.

Or if all gun enthusiasts wanted to band together for some grand Coasian bargain to buy off the potential victims of gun violence, each would have to contribute $220/year to the group effort – not *totally* impossible, but also not something I can really see happening.

This is very, very, very, very *very* tentative, but based on this line of reasoning alone, without looking into the experimental studies or anything else, it appears that Australia-style gun control would probably be worth it, if it were possible.

(I didn't price in the advantages of guns in terms of preventing state tyranny and protecing freedom, which might be worth subsidizing, but my guess is that if 32% gun ownership is enough to maintain freedom, 22% gun ownership is as well)

V.

In summary, with my personal confidence levels:

1. Scatterplots showing raw correlations between gun ownership and "gun deaths" are entirely driven by suicide, and therefore dishonest to use to prove that guns cause murder (~100% confidence)

2. But if you adjust for all relevant confounders, there *is* a positive correlation between gun ownership and homicide rates (~90% confidence). This relationship is likely causal (~66% confidence).

3. The majority of the difference between America's murder rate and that of other First World countries is *not* because of easier access to guns in America (~90% confidence).

4. But some of it *is* due to easier access to guns. This is probably about 0.5 murders/100K/year.

5. An Australian-style gun control program *that worked and had no side effects* would probably prevent about 2,000 murders in the US. It would also prevent a much larger number of suicides. I am otherwise ignoring suicides in this piece because discussing them would make me too angry.

6. Probably the amount of lost gun-related enjoyment an Australian-style gun control program would cause do not outweigh the benefits.

7. This is not really enough analysis to make me have a strong opinion about gun control, since this just looks at the correlational evidence and doesn't really investigate the experimental evidence. Contrary to what everyone always tells you, experimental evidence doesn't always trump correlational – there are cases where each has its strengths – but it wouldn't be responsible to have a *real* opinion on this until I look into that too. Nevertheless, these data are at least *highly consistent with* Australia-style gun control being a good idea for the US.

If you want to look into this more, here is a CSV version of all the relevant data.

Teachers: Much More Than You Wanted To Know

[Epistemic status: This is really complicated, this is not my field, people who have spent their entire lives studying this subject have different opinions, and I don't claim to have done more than a very superficial survey. I welcome corrections on the many inevitable errors.]

I.

Newspapers report that having a better teacher for even a single grade (for example, a better fourth-grade teacher) can improve a child's lifetime earning prospects by $80,000. Meanwhile, behavioral genetics studies suggest that a child's parents have minimal (non-genetic) impact on their future earnings. So one year with your fourth-grade teacher making you learn fractions has vast effects on your prospects, but twenty-odd years with your parents shaping you at every moment doesn't? *Huh?* I decided to try to figure this out by looking into the research on teacher effectiveness more closely.

First, how much do teachers matter compared to other things? To find out, researchers take a district full of kids with varying standardized test scores and try to figure out how much of the variance can be predicted by what school the kids are in, what teacher's class the kids are in, and other demographic factors about the kids. So for example if the test scores of two kids in the same teacher's class were on average no more similar than the test scores of two kids in two different teachers' classes, then teachers can't matter very much. But if we were consistently seeing things like everybody in Teacher A's class getting A+s and everyone in Teacher B's class getting Ds, that would suggest that good teachers are very important.

Here are the results from three teams that tried this (source, source, source):

These differ a little in that the first one assumes away all noise ("unexplained variance") and the latter two keep it in. But they all agree pretty well that individual factors are most important, followed by school and teacher factors of roughly equal size. Teacher factors explain somewhere between 5% and 20% of the variance. Other studies seem to agree, usually a little to the lower end. For example, Goldhaber, Brewer, and Anderson (1999) find teachers explain 9% of variance; Nye, Konstantopoulos, and Hedges (2004) find they explain 13% of variance for math and 7% for reading. The American Statistical Association summarizes the research as "teachers account for about 1% to 14% of the variability in test scores", which seems about right.

So put more simply – on average, individual students' level of ~~ability~~ grit is what makes the difference. Good schools and teachers may push that a little higher, and bad ones bring it a little lower, but they don't work miracles.

(remember that right now we're talking about same-year standardized test scores. That is, we're talking about how much your fourth-grade history teacher affects your performance on a fourth-grade history test. If teacher effects show up anywhere, this is where it's going to be.)

Just as it's much easier to say "this is 40% genetic" than to identify particular genes, so it's much easier to say "this is 10% dependent on school-level factors and 10% based on teacher-level factors" then to identify what those school-level and teacher-level factors are. The Goldhaber study above tries its best, but the only school-level variable they can pin down is that having lots of white kids in your school improves test scores. And as far as I can tell, they don't look at socioeconomic status of the school or its neighborhood, which is probably what the white kids are serving as a proxy for. Even though these "school level effects" are supposed to be things like "the school is well-funded" or "the school has a great principal", I worry that they're capturing student effects by accident. That is, if you go to a school where everyone else is a rich white kid, chances are that means you're a rich white kid yourself. Although they try to control for this, having a couple of quantifiable variables like race and income probably doesn't entirely capture the complexities of neighborhood sorting by social class.

In terms of observable teacher-level effects, the only one they can find that makes a difference is gender (female teachers are better). Teacher certification, years of experience, certification, degrees, et cetera have no effect. This is consistent with most other research, such as Miller, McKenna, and McKenna (1998). A few studies that we'll get to later do suggest teacher experience matters; almost nobody wants to claim certifications or degrees do much.

One measurable variable not mentioned here *does* seem to have a strong ability to predict successful teachers. I'm not able to access these studies directly, but according to the site of the US Assistant Secretary of Education:

> The most robust finding in the research literature is the effect of teacher verbal and cognitive ability on student achievement. Every study that has included a valid measure of teacher verbal or cognitive ability has found that it accounts for more variance in student achievement than any other measured characteristic of teachers (e.g., Greenwald, Hedges, & Lane, 1996; Ferguson & Ladd, 1996; Kain & Singleton, 1996; Ehrenberg & Brewer, 1994).

So far most of this is straightforward and uncontroversial. Teachers account for about 10% of variance in student test scores, it's hard to predict which teachers do better by their characteristics alone, and schools account for a little more but that might be confounded. In order to say more than this we have to have a more precise way of identifying exactly *which* teachers are good, which is going to be more complicated.

II.

Suppose you want to figure out which teachers in a certain district are the best. You know that the only thing truly important in life is standardized test scores [citation needed], so you calculate the average test score for each teacher's class, then crown whoever has the highest average as Teacher Of The Year. What could go wrong?

But you'll probably just give the award to whoever teaches the gifted class. Teachers have classes with very different ability, and we already determined that ~~innate ability~~ grit explains more variance than teacher skill, so teachers who teach disadvantaged children will be at a big, uh, disadvantage.

So okay, back up. Instead of judging teachers by average test score, we can judge them by the average *change* in test score. If they start with a bunch of kids who have always scored around twentieth percentile, and they teach them so much that now the kids score at the fortieth percentile, then even though their kids are still below

average they've clearly done some good work. Rank how many percentile points on average a teacher's students go up or down during the year, and you should be able to identify the best teachers for real this time.

Add like fifty layers of incomprehensible statistics and this is the basic idea behind VAM (value-added modeling), the latest Exciting Educational Trend and the lynchpin of President Obama's educational reforms. If you use VAM to find out which teachers are better than others, you can pay the good ones more to encourage them to stick around. As for the bad ones, VAM opponents are only being slightly unfair when they describe the plan as "firing your way to educational excellence".

A claim like "VAM accurately predicts test scores" is kind of circular, since test scores are what we used to determine VAM. But I *think* the people in this field try to use the VAM of class c to predict the student performance of class c + 1, or other more complicated techniques, and Chetty, Rothstein, and Rivkin, Hanushek, and Kane all find that a one standard deviation increase in teacher VAM corresponds to about a 0.1 standard deviation increase in student test scores.

Let's try putting this in English. Consider an average student with an average teacher. We expect her to score at exactly the 50th percentile on her tests. Now imagine she switched to the best teacher in the whole school. My elementary school had about forty teachers, so this is 97.5th percentile eg two standard deviations above the mean. A teacher whose VAM is two standard deviations above the mean should have students who score on average 0.2 standard deviations above the mean. Instead of scoring at the 50th percentile, now she'll score at the 58th percentile.

Or consider the SAT, which is *not* the sort of standardized test involved in VAM but which at least everybody knows about. Each of its subtests is normed to a mean of 500 and an SD of 110. Our hypothetical well-taught student would go from an SAT of 500 to an SAT of 522. Meanwhile, average SAT subtest score needed to get into Harvard is still somewhere around 740. So this effect is nonzero but not very impressive.

But what happens if we compound this and give this student the best teachers many years in a row? Sanders and Rivers (also Jordan, Mendro, and Weerasinghe) argue the effects are impressive and cumulative. They compare students in Tennessee who got good teachers three years in a row to similar students who got bad teachers three years in a row (good = top quintile; bad = bottom quintile, so only 1/125 students was lucky or unlucky enough to qualify). The average bad-bad-bad student got scores in the 29th percentile; the average good-good-good student got scores in the 83rd percentile – which based on the single-teacher results looks super-additive. This is starting to sound a lot more impressive, and maybe Harvard-worthy after all. In fact, occasionally it is quoted as "four consecutive good teachers would close the black-white achievement gap" (I'm not sure whether this formulation requires also assigning whites to four consecutive bad teachers).

A RAND education report criticizes these studies as "using ad hoc methods" and argue that they're vulnerable to double-counting student achievement. That is, we know that this teacher is the best because her students get great test scores; then later on we return and get excited over the discovery that the best teachers' students get great test scores. Sanders and Rivers did some complicated things that ought to adjust for that; RAND runs simulations and finds that depending on the true size of teacher effects vs. student effects, those complicated things may or may not work. They conclude that "[Sanders and Rivers] provide evidence of the existence and

persistence of teacher or classroom effects, but the size of the effects is likely to be somewhat overstated".

Gary Rubinstein thinks he's debunked Sanders and Rivers style studies. I strongly disagree with his methods – he seems to be saying that the correlation between good teaching and good test scores isn't exactly one and therefore doesn't matter – but he offers some useful data. Just by eyeballing and playing around with it, it looks like most of the gain from these "three consecutive great teachers" actually comes from the last great teacher. So the superadditivity might not be quite right, and Sanders and Rivers might just be genuinely finding bigger teacher effects than anybody else.

At what rate do these gains from good teachers decay?

They decay pretty fast. Jacob, Lefgren and Sims find that only 25% of gains carry on to the next year, and only 15% to the year after that. That is, if you had a great fourth grade teacher who raised your test scores by x points, in fifth grade your test scores will be 0.25x higher than they would otherwise have been. Kane and Rothstein find much the same. A RAND report suggests 20% persistence after one year and 10% persistence after two. Jacob, Lefgren, and Sims find that only 25% of gains remain after one year, and about 13% after two years, after which it drops off much more slowly. All of this contradicts Sanders and Rivers pretty badly.

None of these studies can tell us whether the gains go all the way to zero after a long enough time. Chetty does these calculations and finds that they stabilize at 25% of their original value. But this number is *higher* than the two-year number for most of the other studies, plus Chetty is famous for getting results that are much more spectacular and convenient than anybody else's. I am really skeptical here. I remember a lot more things about last year than I do about twenty years ago, and even though I am pretty sure that my sixth grade teacher (for some weird reason) taught our class line dancing, I can't remember a single dance step. And remember Louis Benezet's early 20th century experiments with not teaching kids any math at all until middle school – after a year or two they were just as good as anyone else, suggesting a dim view of how useful elementary school math teachers must be. And even Chetty doesn't really seem to want to argue the point, saying that his results "[align] with existing evidence that improvements in education raise contemporaneous scores, then fade out in later scores".

In summary, I think there's pretty strong evidence that a +1 SD increase in teacher VAM can increase same-year test scores by + 0.1 SD, but that 50% – 75% of this effect decays in the first two years. I'm less certain how much these numbers change when one gets multiple good or bad teachers in a row, or how fully they decay after the first two years.

III.

When I started looking for evidence about how teachers affected children, I expected teachers' groups and education specialists to be pushing all the positive results. After all, what could be better for them than solid statistical proof that good teachers are super valuable?

In fact, these groups are the strongest *opponents* of the above studies – not because they doubt good teachers have an effect, but because in order to prove that effect you have to concede that good teaching is easy to measure, which tends to turn into proposals to use VAM to measure teacher performance and then fire underperformers.

They argue that VAM is biased and likely to unfairly pull down teachers who get assigned ~~less intelligent~~ lower-grit kids.

It's always fun to watch rancorous academic dramas from the outside, and the drama around VAM is really a level above anything else I've seen. A typical example is the blog VAMboozled! with its oddly hypnotic logo and a steady stream of posts like Kane Is At It Again: "Statistically Significant" Claims Exaggerated To Influence Policy. Historian/researcher Diane Ravitch doesn't have quite as cute an aesthetic, but she writes things like:

> VAM is Junk Science. Looking at children as machine-made widgets and looking at learning solely as standardized test scores may thrill some econometricians, but it has nothing to do with the real world of children, learning, and teaching. It is a grand theory that might net its authors a Nobel Prize for its grandiosity, but it is both meaningless in relation to any genuine concept of education and harmful in its mechanistic and reductive view of humanity.

But tell us how you really feel.

I was originally skeptical of this, but after reading enough of these sites I think they have some good points about how VAM isn't always a good measure.

First, it seems to depend a lot on student characteristics; for example, it's harder to get a high VAM in a class full of English as a Second Language students. It makes perfect sense that ESL students would get low test scores, but since VAM controls for prior achievement you might expect them to get the same VAM anyway. They don't. Also, a lot of VAM models control for student race, gender, socioeconomic status, et cetera. I guess this is better than *not* doing this, but it seems to show a lack of confidence – if controlling for prior achievement was enough, you wouldn't *need* to control for these other things. But apparently people *do* feel the need to control for this stuff, and at that point I bring up my usual objection that you can never control for confounders enough, and also all to some degree these things are probably just lossy proxies for genetics which you *definitely* can't control for enough.

Maybe because of this, there's a lot of noise in VAM estimates. Goldhaber & Hansen (2013) finds that a teacher's VAM in year t is correlated at about 0.3 with their VAM in year t + 1. A Gates Foundation study also found reliabilities from 0.19 to 0.4, averaging about 0.3. Newton et al get slightly higher numbers from 0.4 to 0.6; Bessolo a wider range from 0.2 to 0.6. But these are all in the same ballpark, and Goldhaber and Hanson snarkily note that standardized tests aimed to assess *students* usually need correlations of 0.8 to 0.9 to be considered valid (the SAT, for example, is around 0.87). Although this suggests there's *some* component of VAM which is stable, it can't be considered to be "assessing" teachers in the same way normal tests assess students.

Even if VAM is a very noisy estimate, can't the noise be toned down by averaging it out over many years? I think the answer is yes, and I think the most careful advocates of VAM want to do this, but President Obama wants to improve education *now* and a lot of teachers don't have ten years worth of VAM estimates.

Also, some teachers complain that even averaging it out wouldn't work if there are *consistent* differences in student assignment. For example, if Ms. Andrews always got the best students, and Mr. Brown always got the worst students, then averaging ten years is just going to average ten years of biased data. Proponents argue that aside from a few obvious cases (the teacher of the gifted class, the teacher of the ESL class)

this shouldn't happen. They can add school-fixed effects into their models (eg control for average performance of students at a particular school), leaving behind only teacher effects. And, they argue, which student in a school gets assigned which teacher ought to be random. Opponents argue that it might not be, and cite Paufler and Amrein-Beardsley's survey of principals, in which the principals all admit they don't assign students to classes randomly. But if you look at the study, the principals say that they're trying to be super-random – ie deliberately make sure that all classes are as balanced as possible. Even if they don't 100% achieve this goal, shouldn't the remaining differences be pretty minimal?

Maybe not. Rothstein (2009) tries to "predict" students' *fourth-grade* test scores using their *fifth-grade* teacher's VAM and finds that this totally works. Either schools are defying the laws of time and space, or for some reason the kids who do well in fourth-grade are getting the best fifth-grade teachers. Briggs and Domingue not only replicate these effects, but find that a fifth-grade teacher's "effects" on her students in fourth-grade is *just as big* as her effect on her students when she is actually teaching them, which would suggest that 100% of VAM is bias. Goldhaber has an argument for why there are statistical reasons this might not be so damning, which I unfortunately don't have enough ~~understanding~~ grit to evaluate.

Genetics might also play a role in explaining these results (h/t Spotted Toad's excellent post on the subject). A twin study by Robert Plomin does the classical behavioral genetics thing to VAM and finds that individual students' nth grade VAM is about 40% to 50% heritable. That is, the change in your test scores between third to fourth grade will probably be more like the change in your identical twin's test scores than like the change in your fraternal twin's test scores.

At first glance, this doesn't make sense – since VAM controls for past performance, shouldn't it be a pretty pure measure of your teacher's effectiveness? Toad argues otherwise. One of those Ten Replicated Findings From Behavioral Genetics is that IQ is more shared environmental in younger kids and more genetic in older kids. In other words, when you're really young, how smart you are depends on how enriched your environment is; as you grow older, it becomes more genetically determined.

So suppose that your environment is predisposing you to an IQ of 100, but your genes are predisposing you to an IQ of 120. And suppose (pardon the oversimplification) that at age 5 your IQ is 100, at age 15 it's 120, and change between those ages is linear. Then every year you could expect to gain 2 IQ points. Now suppose there's another kid whose environment is predisposing her to an IQ of 130, but whose genes are predisposing her to an IQ of 90. At age 5 her IQ is 130, at age 15 it's 90, and so every year she is losing 4 IQ points. And finally, suppose that your score on standardized tests is exactly 100% predicted by your IQ. Since you gain two points every year, in fifth grade you'll gain two points on your test, and your teacher will look pretty good. She'll get a good VAM, a raise, and a promotion. Since your friend loses four points every year, in fifth grade she'll lose four points on her test, and her teacher will look incompetent and be assigned remedial training.

This critique meshes nicely with the Rothstein test. Since you're gaining 2 points every year, Prof. Rothstein can use your 5th grade gains of +2 points to accurately predict your fourth grade gain of +2 points. Then he can use your friend's 5th grade loss of -4 points to accurately predict her fourth grade loss of -4 points.

This is a very neat explanation. My only concern is that it doesn't explain decay effects very well. If a fifth grade teacher's time-bending effect on students in fourth

grade is exactly the same as her non-time-bending effect on students in fifth grade, how come her effect on her students once they graduate to sixth grade will only be 25% as large as her fifth grade effects? How come her seventh-grade effects will be smaller still? Somebody here has to be really wrong.

It would be nice to be able to draw all of this together by saying that teachers have almost no persistent effects, and the genetic component identified by Plomin and pointed at by Rothstein represents the 15 – 25% "permanent" gain identified by Chetty and others which so contradicts my lack of line dancing memories. But that would be just throwing out Briggs and Domingue's finding that the Rothstein effect explains 100% of identified VAM.

One thing I kept seeing in the best papers on this was an acknowledgement that instead of arguing "VAMs are biased!" versus "VAMs are great!", people should probably just agree that VAMs are biased, *just like everything else*, and start figuring out ways to measure exactly how biased they are, then use that number to determine what purposes they are or aren't appropriate for. But I haven't seen anybody doing this in a way I can understand.

In summary, there are many reasons to be skeptical of VAM. But some of these reasons contradict each other, and it's not clear that we should be *infinitely* skeptical. A big part of VAM is bias, but there might also be some signal within the noise, especially when it's averaged out over many years.

IV.

So let's go back to that study that says that a good fourth grade teacher can earn you $89,000. The study itself is Chetty, Friedman, and Rockoff (part 1, part 2). You may recognize Chetty as a name that keeps coming up, usually attached to findings about as unbelievable as these ones.

Bloomberg said that "a truly great" teacher could improve a child's earnings by $80,000, but I think this is mostly extrapolation. The number I see in the paper is a claim that a 1 SD better fourth-grade teacher can improve lifetime earnings by $39,000, so let's stick with that.

This sounds impressive, but imagine the average kid works 40 years. That means it's improving yearly earnings by about $1,000. Of note, the study didn't find this. They found that such teachers improved yearly earnings by about $300, but their study population was mostly in their late twenties and not making very much, and they extrapolated that if good teachers could increase the earnings of entry-level workers by $300, eventually they could increase the earnings of workers with a little more experience by $1000. The authors use a lot of statistics to justify this assumption which I'm not qualified to assess. But really, who cares? The fact that having a good fourth grade teacher can improve your adult earnings *any* measurable amount is the weird claim here. Once I accept *that*, I might as well accept $300, $1,000, or $500,000.

And here's the other weird thing. Everyone else has found that teacher effects on test scores decay very quickly over time. Chetty has *sort of* found that up to 25% of them persist, but he doesn't really seem interested in defending that claim and agrees that probably test scores just fade away. Yet as he himself admits, good teachers' impact on earnings works as if there were zero fadeout of teacher effects. He and his co-authors write:

> Our conclusion that teachers have long-lasting impacts may be surprising given evidence that teachers' impacts on test scores "fade out" very rapidly in subsequent grades (Rothstein 2010, Carrell and West 2010, Jacob, Lefgren, and Sims 2010). We confirm this rapid fade-out in our data, but find that teachers' impacts on earnings are similar to what one would predict based on the cross-sectional correlation between earnings and contemporaneous test score gains.

They later go on to call this a "pattern of fade-out and re-emergence", but this is a little misleading. The VAM never re-emerges on test scores. It only shows up in the earnings numbers.

All of this is really dubious, and it seems like Section III gives us an easy way out. There's probably a component of year-to-year stable bias in VAM, such that it captures something about student quality, maybe even innate ability, rather than just teacher quality. It sounds very easy to just say that this is the component producing Chetty's finding of income gains at age 28; students who have higher innate ability in fourth grade will probably still have it in their twenties.

Chetty is aware of this argument and tries to close it off. He conducts a quasi-experiment which he thinks replicates and confirms his original point: what happens when new teachers enter the school?

The thing we're most worried about is bias in student selection to teachers. If we take an entire grade of a school (for example, if a certain school has three fifth-grade teachers, we take all three of them as a unit) this should be immune to such effects. So Chetty looks at entire grades as old teachers retire and new teachers enter. In particular, he looks at such grades when a new teacher transfers from a different school. That new transfer teacher already has a VAM which we know from his work at the other school, which will be either higher or lower than the average VAM of his new school. If it's higher and VAM is real, we should expect the average VAM of that grade of his new school to go up a proportionate amount. If it's lower and VAM is real, we should expect the average VAM of that grade of his new school to go down a proportionate amount. Chetty investigates this with all of the transfer teachers in his data, finds this is in fact what happens, and finds that if he estimates VAM from these transfers he gets the same number (+ $1000 in earnings) that he got from the normal data. This is impressive. Maybe even *too* impressive. Really? The *same* number? So there's *no* bias in the normal data? I thought there was a lot of evidence that *most* of it was bias?

Rothstein is able to replicate Chetty's findings using data from a different district, but then he goes on to do the same thing on Chetty's quasi-experiment as he did on the normal VAMs, with the same results. That is, you can use the amount a school improves when a great new fifth-grade teacher transfers in to predict that teacher's students' fourth-grade performance. Not perfectly. But a little. For some reason, teacher transfers are having the same freaky time-bending effects as other VAM. Rothstein mostly explains this by saying that Chetty incorrectly excluded certain classes and teachers from his sample, although I don't fully understand this argument. He also gives one other example of when this might happen: suppose that a neighborhood is gentrifying. The new teachers who transfer in after the original teachers retire will probably be a better class of professional lured in by the improving neighborhood. And the school's student body will also probably be more genetically and socioeconomically advantaged. So better transfer teachers will be correlated with higher-achieving kids, but they won't have *caused* such high achievement.

After this came an increasingly complicated exchange between Rothstein and Chetty that I wasn't able to follow. Chetty, Friedman, and Rockoff wrote a 52 page Response To Rothstein where they argued that Rothstein's methodology would find retro-causal effects even in a fair experiment where none should exist. According to a 538 article on the debate, a couple of smart people (albeit smart people who already support VAMs and might be biased) think that Chetty's response makes sense, and even Rothstein agrees it "could be" true. 538 definitely thought the advantage in this exchange went to Chetty. But Rothstein responded with a re-replication of his results that he says addresses Chetty's criticisms but still finds the retro-causal effects indicating bias; as far as I know Chetty has not responded and nobody has weighed in to give me an expert opinion on whether or not it's right.

My temptation would usually be to say – here are some really weird results that can't possibly be true which we want to explain away, here's a widely-respected Berkeley professor of economics who says he's explained them away, great, let's forget about the whole thing. But there's one more experiment which I can't dismiss so easily.

V.

Project STAR (Student Teacher Achievement Ratio) was a big educational experiment in the 80s and 90s to see whether or not smaller class size improved student performance. That's a whole different can of worms, but the point is that in order to do this experiment for a while they randomized children to kindergarten classes within schools across 79 different schools. Since one of the biggest possible sources of bias for these last few studies has been possible nonrandom assignment of students to teachers, these Tennessee schools were an opportunity to get much better data than were available anywhere else.

So Chetty, Friedman, Higer, Saez, Schanzenbach, and Yagan analyzed the STAR data. They tried to do a lot of things with predicting earnings based on teacher experience, teacher credentials, and other characteristics, and it's a bit controversial whether they succeeded or not – see Bryan Caplan's analysis (1, 2) for more. Caplan is skeptical of a lot of the study, but one part he didn't address – and which I find most convincing – is based on something a lot like VAM.

Because of the random assignment, Chetty et al don't have to do full VAM here. It looks like their measure of kindergarten teacher quality is just the average of all their students' test scores (wait, kindergarteners are taking standardized tests now? I guess so.) When they're using teacher quality to predict the success of specific students, they use the average of all the test scores *except* that of the student being predicted, in order to keep it fair.

They find that the average test score of all the *other* students in your class, compared against the average score of all the students in other randomly assigned classes in your school, predicts your own test score. "A one percentile increase in entry-year class quality is estimated to raise own test scores by 0.68 percentiles, confirming that test scores are highly correlated across students within a classroom". This fades to approximately zero by fourth grade, confirming that the test-score-related benefits of having a good teacher are transient and decay quickly. *But,* students assigned to a one-percentile-higher class have average earnings that are 0.4% higher at age 25-27! And they say that this relationship is linear! So for example, the best kindergarten teacher in their dataset caused her class to perform at the 70th percentile on average, and these students earned about $17000 on average (remember, these are young entry-level workers in Tennessee) compared to the $15500 or so of their more

average-kindergarten-teacher-having peers. Just their kindergarten teacher, totally apart from any other teacher in their life history, increased their average income 10%. Really, Chetty et al? *Really?*

But as crazy as it is, this study is hard to poke holes in. Even in arguing against it, Caplan notes that "it's an extremely impressive piece" that "the authors are very careful", and that it's "one of the most impressive empirical papers ever written". The experimental randomization means we can't apply most of the usual anti-VAM measures to it. I don't know, man. I just don't know.

Okay, fine. I have one really long-shot possibility. Chetty et al derive their measure for teacher quality from the performance of all of the students in a class, excluding each student in turn as they try to predict his or her results. But this is only exogenous if the student doesn't affect his or her peers' test scores. But it's possible some students *do* affect their peers' test scores. If a student is a behavioral problem, they can screw up the whole rest of their class. Carrell finds that "exposure to a disruptive peer in classes of 25 during elementary school reduces earnings at age 26 by 3 to 4 percent". Now, this in itself is a crazy, hard-to-believe study. But if we accept this second crazy hard-to-believe study, it might provide us with a way of attacking the first crazy hard-to-believe study. Suppose we have a really screwed-up student who is always misbehaving in class and disrupting the lesson. This lowers all his peers' test scores and makes the teacher look low-quality. Then that kid grows up and remains screwed-up and misbehaving and doesn't get as good a job. If this is a big factor in the differences in performances between classes, then so-called "teacher quality" might be conflated with a measure of how many children in their classes are behavioral problems, and apparent effects of teacher quality on earnings might just represent that misbehaving kids tend to become low-earning adults. I'm not sure if the magnitude of this effect checks out, but it *might* be a possibility.

But if we can't make that work, we're stuck believing that good kindergarten teachers can increase your yearly earnings by thousands of dollars. What do we make of that?

Again, everybody finds that test score gains do not last *nearly* that long. So it can't be that kindergarten teachers provide you with a useful fund of knowledge which you build upon later. It can't even be that kindergarten teachers stimulate and enrich you which raises your IQ or makes you love learning or anything like that. It has to be something orthogonal to test scores and measurable intellectual ability.

Chetty et al's explanation is that teachers also teach "non-cognitive skills". I can't understand the regressions they use, but they say that although a one percentile increase in kindergarten class quality has a statistically insignificant increase (+ 0.05 percentiles) on 8th grade test scores, it has a statistically significant increase (+0.15 percentiles) on 8th grade non-cognitive scores ("non-cognitive scores" in this case are a survey where 8th grade teachers answer questions like "does this student annoy others?") They then proceed to demonstrate that the persistence of these non-cognitive effects do a better job of predicting the earning gains than the test scores do. They try to break these non-cognitive effects into four categories: "effort", "initiative", "engagement" and "whether the student values school", but the results are pretty boring and about equally loaded on all of them.

This does go together *really well* with my "behavioral problem" theory of the kindergarten class-earnings effect. The "quality" of a student's kindergarten class, which might have more to do with the number of students who were behavioral problems in it than anything else, doesn't correlate with future test scores but does

correlate with future behavioral problems. It also seems to match Plomin's point about how very early test scores are determined by environment, but later test scores are determined by genetics. A poor learning environment might be a really big deal in kindergarten, but stop mattering as much later on.

But this also goes together with some other studies that have found the same. The test scores gains from pre-K are notorious for vanishing after a couple of years, but a few really big preschool studies like the Perry Preschool Program found that such programs do not boost IQ but may have other effects (though to complicate matters, apparently Perry *did* boost later-life standardized test scores, just not IQ scores, and to further complicate matters, other studies find children who went to pre-K have worse behavior). This also sort of reminds me of some of the very preliminary research I've been linking to recently suggesting that excessively early school starting ages seem to produce an ADHD-like pattern of bad behavior and later-life bad effects, which I was vaguely willing to attribute to overchallenging kids' brains too early while they're still developing. If I wanted to be very mean (and I do!) I could even say that all kindergarten is a neurological insult that destroys later life prospects because of forcing students to overclock their young brains concentrating on boring things, but good teachers can make this less bad than it might otherwise be by making their classes a little more enjoyable.

But even if this is true, it loops back to the question I started with: there's strong evidence that parents have relatively little non-genetic impact on their childrens' life outcomes, but now we're saying that even a kindergarten teacher they only see for a year *does* have such an impact? And what's more, it's not even in the kindergarten teacher's unique area of comparative advantage (teaching academic subjects), but in the domain of *behavioral problems*, something that parents have like one zillion times more exposure to and power over?

I don't know. I still find these studies unbelievable, but don't have the sort of knock-down evidence to dismiss them that I'd like. I'm really impressed with everybody participating in this debate, with the quality of the data, and with the ability to avoid a lot of the usual failure modes. It's just not enough to convince me of anything yet.

VI.

In summary: teacher quality probably explains 10% of the variation in same-year test scores. A +1 SD better teacher might cause a +0.1 SD year-on-year *improvement* in test scores. This decays quickly with time and is probably disappears entirely after four or five years, though there may also be *small* lingering effects. It's hard to rule out the possibility that other factors, like endogenous sorting of students, or students' genetic potential, contributes to this as an artifact, and most people agree that these sorts of scores combine some signal with a lot of noise. For some reason, even though teachers' effects on test scores decay very quickly, studies have shown that they have significant impact on earning as much as 20 or 25 years later, so much so that kindergarten teacher quality can predict thousands of dollars of difference in adult income. This seemingly unbelievable finding has been replicated in quasi-experiments and even in real experiments and is difficult to banish. Since it does not happen through standardized test scores, the most likely explanation is that it involves non-cognitive factors like behavior. I really don't know whether to believe this and right now I say 50-50 odds that this is a real effect or not – mostly based on low priors rather than on any weakness of the studies themselves. I don't understand this field very well and place low confidence in anything I have to say about it.

Further reading: Institute of Education Science summary, Edward Haertel's summary, TTI report, Adler's critique of Chetty, American Statistical Society's critique of Chetty/VAM, Chetty's response, Ballou's critique of Chetty

Antidepressant Pharmacogenomics: Much More Than You Wanted To Know

[Epistemic status: very uncertain. Not to be taken as medical advice. Talk to your doctor before deciding whether or not to get any tests.]

I.

There are many antidepressants in common use. With a few exceptions, none are globally better than any others. The conventional wisdom says patients should keep trying antidepressants until they find one that works for them. If we knew beforehand which antidepressants would work for which patients, it would save everyone a lot of time, money, and misery. This is the allure of pharmacogenomics, the new field of genetically-guided medication prescription.

Everybody has various different types of cytochrome enzymes which metabolize medication. Some of them play major roles in metabolizing antidepressants; usually it's really complicated and several different enzymes can affect the same antidepressant at different stages. But sometimes one or another dominates; for example, Prozac is mostly metabolized by one enzyme called CYP2D6, and Zoloft is mostly metabolized by a different enzyme called CYP2C19.

Suppose (say the pharmacogenomicists) that my individual genetics code for a normal CYP2D6, but a hyperactive CYP2C19 that works ten times faster than usual. Then maybe Prozac would work normally for me, but every drop of Zoloft would get shredded by my enzymes before it can even get to my brain. A genetic test could tell my psychiatrist this, and then she would know to give me Prozac and not Zoloft. Some tests like this are already commercially available. Preliminary results look encouraging. As always, the key words are "preliminary" and "look", and did I mention that these results were mostly produced by pharma companies pushing their products?

But let me dream for a just a second. There's been this uneasy tension in psychopharmacology. Clinical psychiatrists give their patients antidepressants and see them get better. Then research psychiatrists do studies and show that antidepressant effect sizes are so small as to be practically unnoticeable. The clinicians say "Something must be wrong with your studies, we see our patients on antidepressants get much better all the time". The researchers counter with "The plural of anecdote isn't 'data', your intuitions deceive you, antidepressant effects are almost imperceptibly weak." At this point we prescribe antidepressants anyway, because – what else are you going to do when someone comes into your office in tears and begs for help? – but we feel kind of bad about it.

Pharmacogenomics offers a way out of this conundrum. Suppose half of the time patients get antidepressants, their enzymes shred the medicine before it can even get to the brain, and there's no effect. In the other half, the patients have normal enzymes, the medications reach the brain, and the patient gets better. Researchers would average together all these patients and conclude "Antidepressants have an effect, but on average it's very small". Clinicians would keep the patients who get good effects, keep switching drugs for the patients who get bad effects until they find something that works, and say "Eventually, most of my patients seem to have good effects from antidepressants".

There's a little bit of support for this in studies. STAR*D found that only 33% of patients improved on their first antidepressant, but that if you kept changing antidepressants, about 66% of patients would eventually find one that helped them improve. Gueorguieva & Mallinckrodt (2011) find something similar by modelling "growth trajectories" of antidepressants in previous studies. If it were true, it would be a big relief for everybody.

It might also mean that pharmacogenomic testing would solve the whole problem forever and lets everyone be on an antidepressant that works well for them. Such is the dream.

But pharmacogenomics still very young. And due to a complicated series of legal loopholes, it isn't regulated by the FDA. I'm mostly in favor of more things avoiding FDA regulation, but it means the rest of us have to be much more vigilant.

A few days ago I got to talk to a representative of the company that makes GeneSight, the biggest name in pharmacogenomic testing. They sell a $2000 test which analyzes seven genes, then produces a report on which psychotropic medications you might do best or worst on. It's exactly the sort of thing that would be great if it worked – so let's look at it in more depth.

II.

GeneSight tests seven genes. Five are cytochrome enzymes like the ones discussed above. The other two are HTR2A, a serotonin receptor, and SLC6A4, a serotonin transporter. These are obvious and reasonable targets if you're worried about serotonergic drugs. But is there evidence that they predict medication response?

GeneSight looks at the rs6313 SNP in HTR2A, which they say determines "side effects". I *think* they're thinking of Murphy et al (2003), who found that patients with the (C,C) genotype had worse side effects on Paxil. The study followed 122 patients on Paxil, of whom 41 were (C,C) and 81 were something else. 46% of the (C,C) patients hated Paxil so much they stopped taking it, compared to only 16% of the others ($p = 0.001$). There was no similar effect on a nonserotonergic drug, Remeron. This study is interesting, but it's small and it's never been replicated. The closest thing to replication is this study which focused on nausea, the most common Paxil side effect; it found the gene had no effect. This study looked at Prozac and found that the gene didn't affect Prozac response, but it didn't look at side effects and didn't explain how it handled dropouts from the study. I am really surprised they're including a gene here based on a small study from fifteen years ago that was never replicated.

They also look at SLC6A4, specifically the difference between the "long" versus "short" allele. This has been studied *ad nauseum* – which isn't to say anyone has come to any conclusions. According to Fabbri, Di Girolamo, & Serretti, there are 25 studies saying the long allele of the gene is better, 9 studies saying the short allele is better, and 20 studies showing no difference. Two meta-analyses (1 n = 1435, 2 n = 5479) come out in favor of the long allele; two others (1 n = 4309, 2, n = 1914) fail to find any effect. But even the people who find the effect admit it's pretty small – the Italian group estimates 3.2%. This would both explain why so many people miss it, and relieve us of the burden of caring about it at all.

The Carlat Report has a conspiracy theory that GeneSight really only uses the liver enzyme genes, but they add in a few serotonin-related genes so they can look cool; presumably there's more of a "wow" factor in directly understanding the target receptors in the brain than in mucking around with liver enzymes. I like this theory.

Certainly the results on both these genes are small enough and weak enough that it would be weird to make a commercial test out of them. The liver enzymes seem to be where it's at. Let's move on to those.

The Italian group that did the pharmacogenomics review mentioned above are not sanguine about liver enzymes. They write (as of 2012, presumably based on Genetic Polymorphisms Of Cytochrome P450 Enzymes And Antidepressant Metabolism">this previous review):

> Available data do not support a correlation between antidepressant plasma levels and response for most antidepressants (with the exception of TCAs) and this is probably linked to the lack of association between response and CYP450 genetic polymorphisms found by the most part of previous studies. In all facts, the first CYP2D6 and CYP2C19 genotyping test (AmpliChip) approved by the Food and Drug Administration has not been recommended by guidelines because of lack of evidence linking this test to clinical outcomes and cost-effectiveness studies.

What does it even mean to say that there's no relationship between SSRI plasma level and therapeutic effect? Doesn't the drug only work when it's in your body? And shouldn't the amount in your body determine the effective dose? The only people I've found who even begin to answer this question are Papakostas & Fava, who say that there are complicated individual factors determining how much SSRI makes it from the plasma to the CNS, and how much of it binds to the serotonin transporter versus other stuff. This would be a lot more reassuring if amount of SSRI bound to the serotonin transporter correlated with clinical effects, which studies seem very uncertain about. I'm not really sure how to fit this together with SSRIs having a dose-dependent effect, and I worry that somebody must be very confused. But taking all of this at face value, it doesn't really look good for using cytochrome enzymes predicting response.

I talked to the GeneSight rep about this, and he agreed; their internal tests don't show strong effects for any of the candidate genes alone, because they all interact with each other in complicated ways. It's only when you look at all of them together, using the proprietary algorithm based off of their proprietary panel, that everything starts to come together.

This is possible, but given the poor results of everyone else in the field I think we should take it with a grain of salt.

III.

We might also want to zoom out and take a broader picture: should we *expect* these genes to matter?

It's much easier to find the total effect of genetics than it is to find the effect of any individual gene; this is the principle behind twin studies and GCTAs. Tansey et al do a GCTA on antidepressant response and find that all the genetic variants tested, combined, explain 42% of individual differences in antidepressant response. Their methodology allowed them to break it down chromosome-by-chromosome, and they found that genetic effects were pretty evenly distributed across chromosomes, with longer chromosomes counting more. This is consistent with massively polygenic structure where there are hundreds of thousands of genes, each of small effects – much like height or IQ. But typically even the strongest IQ or height genes only explain about 1% of the variance. So an antidepressant response test containing only seven genes isn't likely to do very much *even if those genes are correctly chosen and well-understood*.

SLC6A4 is a great example of this. It's on chromosome 17. According to Tansey, chromosome 17 explains less than 1% of variance in antidepressant effect. So unless Tansey is very wrong, SLC6A4 must also explain less than 1% of the variance, which means it's clinically useless. The other six genes on the test aren't looking great either.

Does this mean that the GeneSight panel *must* be useless? I'm not sure. For one thing, the genetic structure of *which* antidepressant you respond to might be different from the structure of antidepressant response generally (though the study found similar structures to any-antidepressant response and SSRI-only response). For another, for complicated reasons sometimes *exploiting* variance is easier than *predicting* variance; I don't understand this enough to be sure that this isn't one of these cases, though it doesn't look that way to me.

I don't think this is a knock-down argument against anything. But I think it means we should take any claims that a seven (or ten, or fifty) gene panel can predict very much with *another* grain of salt.

IV.

But assuming that there are relatively few genes, and we figure out what they are, then we're basically good, right? Wrong.

Warfarin is a drug used to prevent blood clots. It's notorious among doctors for being finicky, confusing, difficult to dose, and making people to bleed to death if you get it wrong. This made it a very promising candidate for pharmacogenomics: what if we could predict everyone's individualized optimal warfarin dose and take out the guesswork?

Early efforts showed promise. Much of the variability was traced to two genes, VKORC1 and CYP2C9. Companies created pharmacogenomic panels that could predict warfarin levels pretty well based off of those genes. Doctors were urged to set warfarin doses based on the results. Some initial studies looked positive. Caraco et al and Primohamed et al both found in randomized controlled trials with decent sample sizes that warfarin patients did better on the genetically-guided algorithm, $p < 0.001$. A 2014 meta-analysis looked at nine studies of the algorithm, over 2812 patients, and found that it didn't work. Whether you used the genetic test or not didn't affect number of blood clots, percent chance of having your blood within normal clotting parameters, or likelihood of major bleeding. There wasn't even a marginally significant trend. Another 2015 meta-analysis found the same thing. Confusingly, a Chinese group did a third meta-analysis that *did* find advantages in some areas, but Chinese studies tend to use shady research practices, and besides, it's two to one.

UpToDate, the canonical medical evidence aggregation site for doctors, concludes:

> We suggest not using pharmacogenomic testing (ie, genotyping for polymorphisms that affect metabolism of warfarin and vitamin K-dependent coagulation factors) to guide initial dosing of the vitamin K antagonists (VKAs). Two meta-analyses of randomized trials (both involving approximately 3000 patients) found that dosing incorporating hepatic cytochrome P-450 2C9 (CYP2C9) or vitamin K epoxide reductase complex (VKORC1) genotype did not reduce rates of bleeding or thromboembolism.

I mention this to add another grain of salt. Warfarin is the perfect candidate for pharmacogenomics. It's got a lot of really complicated interpersonal variation that

often leads to disaster. We know this is due to only a few genes, and we know exactly which genes they are. We understand pretty much every aspect of its chemistry perfectly. Preliminary studies showed amazing effects.

And yet pharmacogenomic testing for warfarin basically doesn't work. There are a few special cases where it can be helpful, and I think the guidelines say something like "if you have your patient's genotype already for some reason, you might as well use it". But overall the promise has failed to pan out.

Antidepressants are in a worse place than warfarin. We have only a vague idea how they work, only a vague idea what genes are involved, and plasma levels don't even consistently correlate with function. It would be very strange if antidepressant testing worked where warfarin testing failed. But, of course, it's not impossible, so let's keep our grains of salt and keep going.

V.

Why didn't the warfarin pharmacogenomics work? They had the genes right, didn't they?

I'm not too sure what's going on, but maybe it just didn't work better than doctors titrating the dose the old-fashioned way. Warfarin is a blood thinner. You can take blood and check how thin it is, usually measured with a number called INR. Most warfarin users are aiming for an INR between 2 and 3. So suppose (to oversimplify) you give your patient a dose of 3 mg, and find that the INR is 1.7. It seems like maybe the patient needs a little more warfarin, so you increase the dose to 4 mg. You take the INR later and it's 2.3, so you declare victory and move on.

Maybe if you had a high-tech genetic test you could read the microscopic letters of the code of life itself, run the results through a supercomputer, and determine from the outset that 4 mg was the optimal dose. But all it would do is save you a little time.

There's something similar going on with depression. Starting dose of Prozac is supposedly 20 mg, but I sometimes start it as low as 10 to make sure people won't have side effects. And maximum dose is 80 mg. So there's almost an order of magnitude between the highest and lowest Prozac doses. Most people stay on 20 to 40, and that dose seems to work pretty well.

Suppose I have a patient with a mutation that slows down their metabolism of Prozac; they effectively get three times the dose I would expect. I start them on 10 mg, which to them is 30 mg, and they seem to be doing well. I increase to 20, which to them is 60, and they get a lot of side effects, so I back down to 10 mg. Now they're on their equivalent of the optimal dose. How is this worse than a genetic test which warns me against using Prozac because they have mutant Prozac metabolism?

Or suppose I have a patient with a mutation that *dectuples* Prozac levels; now there's *no* safe dose. I start them on 10 mg, and they immediately report terrible side effects. I say "Yikes", stop the Prozac, and put them on Zoloft, which works fine. How is this worse than a genetic test which says Prozac is bad for this patient but Zoloft is good?

Or suppose I have a patient with a mutation that makes them an ultrarapid metabolizer; no matter how much Prozac I give them, zero percent ever reaches their brain. I start them on Prozac 10 mg, nothing happens, go up to 20, then 40, then 60, then 80, nothing happens, finally I say "Screw this" and switch them to Zoloft. Once again, how is this worse than the genetic test?

(again, all of this is pretending that dose correlates with plasma levels correlates with efficacy in a way that's hard to prove, but presumably necessary for any of this to be meaningful at all)

I expect the last two situations to be very rare; few people have orders-of-magnitude differences in metabolism compared to the general population. Mostly it's going to be people who I would expect to need 20 of Prozac actually needing 40, or vice versa. But nobody has the slightest idea how to dose SSRIs anyway and we usually just try every possible dose and stick with the one that works. So I'm confused how genetic testing is supposed to make people do better or worse, as opposed to just needing a little more or less of a medication whose dosing is so mysterious that nobody ever knows how much anyone needs anyway.

As far as I can tell, this is why they need those pharmacodynamic genes like HTR2A and SLC6A4. Those represent real differences between antidepressants and not just changes in dose which we would get to anyway. I mean, you could still just switch antidepressants if your first one doesn't work. But this would admittedly be hard and some people might not do it. *Everyone* titrates doses!

This is a fourth grain of salt and another reason why I'm wary about this idea.

VI.

Despite my skepticism, there are several studies showing impressive effects from pharmacogenomic antidepressant tests. Now that we've established some reasons to be doubtful, let's look at them more closely.

GeneSight lists eight studies on its website here. Of note, all eight were conducted by GeneSight; as far as I know no external group has ever independently replicated any of their claims. The GeneSight rep I talked to said they're trying to get other scientists to look at it but haven't been able to so far. That's fair, but it's also fair for me to point out that studies by pharma companies are far more likely to find their products effective than studies by anyone else (OR = 4.05). I'm not going to start a whole other section for this, but let's call it a fifth grain of salt.

First is the LaCrosse Clinical Study. 114 depressed patients being treated at a clinic in Wisconsin received the GeneSight test, and the results were given to their psychiatrists, who presumably changed medications in accordance with the tests. Another 113 depressed patients got normal treatment without any genetic testing. The results were:

Taken from here, where you'll find much more along the same lines.

All of the combinations of letters and numbers are different depression tests. The blue bars are the people who got genotyped. The grey bars are the people who didn't. So we see that on every test, the people who got genotyped saw much greater improvement than the people who didn't. The difference in remission was similarly impressive; by 8 weeks, 26% of the genotyped group were depression-free as per QIDS-C16 compared to only 13% of the control group (p = 0.03)

How can we nitpick these results? A couple of things come to mind.

Number one, the study wasn't blinded. Everyone who was genotyped knew they were genotyped. Everyone who wasn't genotyped knew they weren't genotyped. I'm still not sure whether there's a significant placebo effect in depression (Hróbjartsson and Gøtzsche say no!), but it's at least worth worrying about.

Number two, the groups weren't randomized. I have no idea why they didn't randomize the groups, but they didn't. The first hundred-odd people to come in got put in the control group. The second hundred-off people got put in the genotype group. In accordance with the prophecy, there are various confusing and inexplicable differences between the two groups. The control group had more previous medication trials (4.7 vs. 3.6, p = 0.02). The intervention group had higher QIDS scores at baseline (16 vs. 17.5, p = 0.003). They even had different CYP2D6 phenotypes (p = 0.03). On their own these differences don't seem so bad, but they raise the question of why these groups were different at all and what other differences might be lurking.

Number three, the groups had very different numbers of dropouts. 42 people dropped out of the genotyped group, compared to 20 people from the control group. Dropouts made up about a quarter of the entire study population. The authors theorize that people were more likely to drop out of the genotype group than the control group because they'd promised to give the control group their genotypes at the end of the study, so they were sticking around to get their reward. But this means that people who were failing treatment were likely to drop out of the genotype group (making them look better) but stay in the control group (making them look worse). The authors do an analysis and say that this didn't affect things, but it's another crack in the study.

All of these are bad, but intuitively I don't feel like any of them should have been able to produce as dramatic an effect as they actually found. But I do have one theory about how this might have happened. Remember, these are all people who are on antidepressants already but aren't getting better. The intervention group's doctors get genetic testing results saying what antidepressant is best for them; the control group's doctors get nothing. So the intervention group's doctors will probably switch their patients' medication to the one the test says will be best, and the control group's doctors might just leave them on the antidepressant that's already not working. Indeed, we find that 77% of intervention group patients switched medications, compared to 44% of control group patients. So imagine if the genetic test didn't work at all. 77% of intervention group patients at least switch off their antidepressant that *definitely* doesn't work and onto one that *might* work; meanwhile, the control group mostly stays on the same old failed drugs.

Someone (maybe Carlat again?) mentioned how they should have controlled this study: give everyone a genetic test. Give the intervention group their own test results, and give the control group *someone else's* test results. If people do better on their own results than on random results, *then* we're getting somewhere.

Second is the Hamm Study, which is so similar to the above I'm not going to treat it separately.

Third is the Pine Rest Study. This one is, at least, randomized and *single*-blind. Single-blind means that the patients don't know which group they're in, but their doctors do; this is considered worse than double-blind (where neither patients nor doctors know) because the doctors' subtle expectations could unconsciously influence the patients. But at least it's something.

Unfortunately, the sample size was only 51 people, and the p-value for the main outcome was 0.28. They tried to salvage this with some subgroup analyses, but f**k that.

Fourth and fifth are two different meta-analyses of the above three studies, which is the lowest study-to-meta-analysis ratio I've ever seen. They find big effects, but "garbage in, garbage out".

Sixth, there's the Medco Study by Winner et al; I assume his name is a Big Pharma plot to make us associate positive feelings with him. This study is an attempt to prove cost-effectiveness. The GeneSight test costs $2000, but it might be worth it to insurers/governments if it makes people so much healthier that they spend less money on health care later. And indeed, it finds that GeneSight users spend $1036 less per year on medication than matched controls.

The details: they search health insurance databases for patients who were taking an psychiatric medication and then got GeneSight tests. Then they search the same databases for control patients for each; the control patients take the same psych med, have the same gender, are similar in age, and have the same primary psychiatric diagnosis. They end up with 2000 GeneSight patients and 10000 matched controls, whom they prove are definitely similar (even as a group) on the traits mentioned above. Then they follow all these people for a year and see how their medication spending changes.

The year of the study, the GeneSight patients spent on average $689 more on medications than they did the year before – unfortunate, but not entirely unexpected since apparently they're pretty sick. The control patients spent on average $1725 more. So their medication costs increased much more than the GeneSight patients. That presumably suggests GeneSight was doing a good job treating their depression, thus keeping costs down.

The problem is, this study wasn't randomized and so I see no reason to expect these groups to be comparable in any way. The groups were matched for sex, age, diagnosis, and one drug, but not on any other basis. And we have reason to think that they're not the same – after all, one group consists of people who ordered a little-known $2000 genetic test. To me, that means they're probably 1) rich, and 2) have psychiatrists who are really cutting-edge and into this kind of stuff. To be fair, I would expect both of those to drive *up* their costs, whereas in fact their costs were lower. But consider the possibility that rich people with good psychiatrists probably have less severe disease and are more likely to recover.

Here's some more evidence for this: of the ~$1000 cost savings, $300 was in psychiatric drugs and $700 was in non-psychiatric drugs. The article mentions that there's a mind-body connection and so maybe treating depression effectively will make people's non-psychiatric diseases get better too. This is *true*, but I think seeing that the effect of a psychiatric intervention is *stronger* on non-psychiatric than psychiatric conditions should at least *raise our suspicion* that we're actually seeing some confounder.

I cannot find anywhere in the study a comparison of how much money each group spent the year before the study started. This is a very strange omission. If these numbers were very different, that would clinch this argument.

Seventh is the Union Health Service study. They genotype people at a health insurance company who have already been taking a psychotropic medication. The

genetic test either says that their existing medication is good for them ("green bin"), okay for them ("yellow bin") or bad for them ("red bin"). Then they compare how the green vs. yellow vs. red patients have been doing over the past year on their medications. They find green and yellow patients mostly doing the same, but red patients doing very badly; for example, green patients have about five sick days from work a year, but red patients have about twenty.

I don't really see any obvious flaws in this study, but there are only nine red patients, which means their entire results depend on an n = 9 experimental group.

Eighth is a study that just seems to be a simulation of how QALYs might change if you enter some parameters; it doesn't contain any new empirical data.

Overall these studies show very impressive effects. While it's possible to nitpick all of them, we have to remind ourselves that we can nitpick anything, even the best of studies, and do we really want to be that much of a jerk when these people have tested their revolutionary new product in five different ways, and every time it's passed with flying colors aside from a few minor quibbles?

And the answer is: yes, I want to be *exactly* that much of a jerk. The history of modern medicine is one of pharmaceutical companies having *amazing* studies supporting their product, and maybe if you squint you can just *barely* find one or two little flaws but it hardly seems worth worrying about, and then a few years later it comes out that the product had no benefits whatsoever and caused everyone who took it to bleed to death. The reason for all those grains of salt above was to suppress our natural instincts toward mercy and cultivate the proper instincts to use when faced with pharmaceutical company studies, ie Cartesian doubt mixed with smoldering hatred.

VII.

I am totally not above introducing arguments from authority, and I've seen two people with much more credibility than myself look into this. The first is Daniel Carlat, Tufts professor and editor of *The Carlat Report*, a well-respected newsletter/magazine for psychiatrists. He writes a skeptical review of their studies, and finishes:

> If we were to hold the GeneSight test to the usual standards we require for making medication decisions, we'd conclude that there's very little reliable evidence that it works.

The second is John Ioannidis, professor of health research at Stanford and universally recognized expert on clinical evidence. He doesn't look at GeneSight in particular, but he writes of the whole pharmacogenomic project:

> For at least 3 years now, the expectation has been that newer platforms using exome or full-genome sequencing may improve the genome coverage and identify far more variants that regulate phenotypes of interest, including pharmacogenomic ones. Despite an intensive research investment, these promises have not yet materialized as of early 2013. A PubMed search on May 12, 2013, with (pharmacogenomics* OR pharmacogenetc*) AND sequencing yielded an impressive number of 604 items. I scrutinized the 80 most recently indexed ones. The majority were either reviews/commentary articles with highly promising (if not zealot) titles or irrelevant articles. There was not a single paper that had shown robust statistical association between a newly discovered gene and some pharmacogenomics outcome, detected by sequencing. If anything, the few articles with real data, rather than promises, show that the task of detecting and

> validating statistically rigorous associations for rare variants is likely to be formidable. One comprehensive study sequencing 202 genes encoding drug targets in 14,002 individuals found an abundance of rare variants, with 1 rare variant appearing every 17 bases, and there was also geographic localization and heterogeneity. Although this is an embarrassment of riches, eventually finding which of these thousands of rare variants are most relevant to treatment response and treatment-related harm will be a tough puzzle to solve even with large sample sizes.
>
> Despite these disappointing results, the prospect of applying pharmacogenomics in clinical care has not abided. If anything, it is pursued with continued enthusiasm among believers. But how much of that information is valid and is making any impact? […]
>
> Before investing into expensive clinical trials for testing the new crop of mostly weak pharmacogenomic markers, a more radical decision is whether we should find some means to improve the yield of pharmacogenomics or just call it a day and largely abandon the field. The latter option sounds like a painfully radical solution, but on the other hand, we have already spent many thousands of papers and enormous funding, and the yield is so minimal. The utility yield seems to be even diminishing, if anything, as we develop more sophisticated genetic measurement techniques. Perhaps we should acknowledge that pharmacogenomics was a brilliant idea, we have learned some interesting facts to date, and we also found a handful of potentially useful markers, but industrial-level application of research funds may need to shift elsewhere.

I think the warning from respected authorities like these should add a sixth grain of salt to our rapidly-growing pile and make us feel a little bit better about rejecting the evidence above and deciding to wait.

VIII.

There's a thing I always used to hate about the skeptic community. Some otherwise-responsible scientist would decide to study homeopathy for some reason, and to everyone's surprise they would get positive results. And we would be uneasy, and turn to the skeptic community for advice. And they would say "Yeah, but homeopathy is stupid, so forget about this." And they would be *right*, but – what's the point of having evidence if you ignore it when it goes the wrong way? And what's the point in having experts if all they can do is say "this evidence went the wrong way, so let's ignore it"? Shouldn't we demand experts so confident in their understanding that they can explain to us why the new "evidence" is wrong? And as a corollary, shouldn't we demand experts who – if the world really was topsy-turvy and some crazy alternative medicine scheme did work – would be able to recognize that and tell us when to suspend our usual skepticism?

But at this point I'm starting to feel a deep kinship with skeptic bloggers. Sometimes we can figure out *possible* cracks in studies, and I think Part VI above did okay with that. But there will be cracks in even the best studies, and there will *especially* be cracks in studies done by small pharmaceutical companies who don't have the resources to do a major multicenter trial, and it's never clear when to use them as an excuse to reject the whole edifice versus when to let them pass as an unavoidable part of life. And because of how tough pharmacogenomics has proven so far, this is a case where I – after reading the warnings from Carlat and Ioannidis and the Italian team and everyone else – tentatively reject the edifice.

I hope later I kick myself over this. This *might* be the start of a revolutionary exciting new era in psychiatry. But I don't think I can believe it until independent groups have evaluated the tests, until other independent groups have replicated the work of the first independent groups, until everyone involved has publicly released their data (GeneSight didn't release any of the raw data for any of these studies!), and until our priors have been raised by equivalent success in other areas of pharmacogenomics.

Until then, I think it is a neat toy. I am glad some people are studying it. But I would not recommend spending your money on it if you don't have $2000 to burn (though I understand most people find ways to make their insurance or the government pay).

But if you just want to have fun with this, you can get a cheap approximation from 23andMe. Use the procedure outlined here to get your raw data, then look up rs6313 for the HTR2A polymorphism; (G,G) supposedly means more Paxil side effects (and maybe SSRI side effects in general). 23andMe completely dropped the ball on SLC6A4 and I would not recommend trying to look that one up. The cytochromes are much more complicated, but you might be able to piece some of it together from this page's links links to lists of alleles and related SNPs for each individual enzyme; also Promethease will do some of it for you automatically. Right now I think this process would produce pretty much 100% noise and be completely useless. But I'm not sure it would be *more* useless than the $2000 test. And if any of this pharmacogenomic stuff turns out to work, I hope some hobbyist automates the 23andMe-checking process and sells it as shareware for $5.

A Story With Zombies

(inspired by Zombies: Seriously, Enough, Zombies Are So Overdone, and Scifi/Fantasy Stories Editors Are Tired Of Seeing: Zombies)

He walked into my office and threw the manuscript on my desk with a thud.

“It’s called *Thankful For Zombies*. A zombie story where…”

“Nope,” I said.

His face deflated like a balloon. “But I didn’t even…”

“Zombies are overdone,” I said.

“But this is a zombie story with a twist!”

“Zombie stories with twists are *super* overdone.”

“But this is a story about an extended family who get together for Thanksgiving dinner, only to be interrupted by a zombie apocalypse. It’s a Thanksgiving story about zombies. You have to admit that the combination of zombies and Thanksgiving has never…”

“Done,” I said.

“Wait, really? The family starts out estranged and suspicious of each other, but then when they all have to work together to…”

“Done,” I said.

“How could that have been done?”

“Listen. I know you won’t believe me, but for the past ten years or so, the best literary minds of our generation have been working on creating zombie stories *just* different enough from every other zombie story around to get published. First the clever and interesting twists got explored. Then the mediocre and boring twists. Then the absurd and idiotic twists. Finally the genre got *entirely mined out*. There is now a New York Times bestselling book about zombies invading Jane Austen’s *Pride and Prejudice*. If your idea isn’t weirder than that, *it’s been done*. And that’s the logical ‘if’. If your idea *is* weirder than that, *it has also been done*.”

“I *will* get *Thankful for Zombies* published,” he said.

“You won’t,” I advised him.

“I just have to think of an original angle.”

“You really won’t,” I told him.

“The zombies are the good guys,” he proposed.

“Done.”

“The zombies are smarter than humans.”

“Done.”

“In the end, we ourselves are the zombies.”

“Done.”

“A human girl falls in love with a zombie.”

“Done.”

“Okay, fine. Toss the Thanksgiving angle. There’s got to be some zombie plot that will be fresh and new.”

“I promise you, there’s not.”

“Zombies in space.”

“Done.”

“Zombies *from* space.”

“Done.”

“Zombies *are* space.”

“Done.”

“Zombies in Victorian England.”

“Done‘”

“Zombies in Edwardian England.”

“Done.”

“Zombies in Shakespearean England.”

“Done.”

“Shakespeare was a zombie, and all of his plays are just the word BRAAAAAAIIINS repeated over and over again.”

“Done, for some reason.”

“A young zombie comes of age.”

“Done.”

“A middle-aged zombie wonders if her single-minded focus on career success has prevented her from becoming the kind of zombie she wanted to be when she was younger.”

“Done.”

“An elderly zombie contemplates death.”

“Zombies are already dead.”

“Then I can…”

“…and yet it’s still been done.”

“A zombie in the Vietnam War.”

“Done.”

“A hippie zombie at Woodstock.”

“Done.”

“Strong female zombies.”

“Done.”

“Jewish zombies.”

“Done.”

“Black zombies.”

“Done.”

“A gay zombie struggling to fit into a homophobic zombie society.”

“Come on, this is the 21st century. Done like ten times. One of them won the Booker.”

“Gender-questioning zombies.”

“Done.”

“An immigrant zombie comes to America, with nothing but the decaying shirt on his back, knowing only a single word of English.”

“*All* zombies only know a single word of English. Also, done.”

“Nazi zombies.”

“Done.”

“Vampire zombies.”

“Done.”

“Pirate zombies.”

“Done.”

“Obstetrician/gynaecologist zombies.”

“Done.”

“Zombie Hitler.”

“Done.”

“Zombie Henry VIII.”

“Done“.

“But what if it was told from the perspective of Anne Boleyn?”

“Done.”

“Zombie Leonardo da Vinci.”

“Done.”

“Zombie Jesus.”

“Done. By three guys named Matt, Luke, and John.”

“Zombie Buddha.”

“Done.”

“Zombie Mohammed.”

“Done. As is the author, if you get my drift.”

“Zombie Zoroaster.”

“Done.”

“A parody subverting zombie stories.”

“Super done.”

“A parody subverting zombie stories lampshading how overdone they are.”

“Super duper done.”

“Hmmmm.” He thinks for a second. “Hold on, I’m remembering something from my college math class that might work here. You take all the zombie novels ever written, and you put them in some well-ordering, for example from first to last published. Then you make a new novel, consisting of the first page of the first novel, the second page of the second novel, and so on. But you change each page just a little bit. Since we know the first page of the new novel is different from the first page of the first novel, and the second page of the new novel is different from the second page of the second novel, by extension we know that there is at least one page on which the new novel is different from each zombie novel currently in existence. That means that the new story is provably original.”

“Done.”

“I don’t think you understand; it’s mathematically impossible for…”

“No, I mean there’s a story about a zombie doing that.”

“Oh.” He furrowed his brow. “A zombie superhero.”

“Done.”

“Steampunk zombies.”

“Done. I think now you’re just trolling me.”

“Motorcycle gangs of zombies.”

“Done.”

“A zombie story that’s a metaphor for how…”

“Done.”

“I didn’t finish!”

“You didn’t have to.”

“A zombie gets cancer.”

“Done.”

“A zombie gets depression.”

“Done.”

“A zombie tries to write zombie fiction.”

“Done.”

“A zombie tries to write zombie fiction *about* a zombie trying to write zombie fiction.”

“Done.”

“A zombie tries to…”

“It’s done all the way down.”

“Young free-spirited zombies trying to see America.”

“Done.”

“A story that starts off as being about a fantasy society of knights and damsels, but at the very end it’s revealed everyone is a zombie.”

“Done.”

“A story that starts off as being about a young woman’s struggle to succeed in 1980s Wall Street, but at the very end it’s revealed everyone is a zombie.”

“Done.”

“A story that starts off as being a paleontology textbook about the fauna of the Lower Cretaceous, but at the very end it’s revealed everyone is a zombie.”

“Twist zombie endings are *done*.”

“A zombie…a zombie riding a giant purple emu through 17th century Ireland teams up with the pre-ghost of Thomas Jefferson to investigate a crime in which time-traveling flamboyantly gay sapient hippos have murdered the Secret Protestant Pope in order to initiate the Jain apocalypse, with liberal quotations from and allusions to the works of Edgar Allen Poe Thomas Pynchon and the medieval Rolandic cycle, and also the whole thing is a metaphor for Republican resistance to climate change legislation.”

I thought for a moment. “Okay,” I said. “That particular plot has not, technically, been done. But no one would read it.”

“They will,” he said.

“You’d be wasting your time to write it.”

“I’m writing it,” he said.

“Suit yourself. Put it on my desk when you’re finished, and I’ll take a look at it. But your chances aren’t good.”

“I don’t care,” he said, and left.

I sighed, finished up my last couple of pieces of paperwork, and shambled home from the office. On the way out, I ate my secretary’s brain.

Asches to Asches

[Content note: fictional story contains gaslighting-type elements. May induce Cartesian skepticism]

You wake up in one of those pod things like in *The Matrix*. There's a woman standing in front of you, wearing a lab coat, holding a clipboard.

"Hi," she says. "This is the real world. You used to live here. We erased your memories and stuck you in a simulated world for a while, like in *The Matrix*. It was part of a great experiment."

"What?" you shout. "My whole life, a lie? How dare you deceive me as part of some grand 'experiment' I never consented to?"

"Oh," said the woman, "actually, you did consent, in exchange for extra credit in your undergraduate psychology course." She hands you the clipboard. There is a consent form with your name on it, in your handwriting.

You give her a sheepish look. "What was the experiment?"

"You know families?" asks the woman.

"Of course," you say.

"Yeah," says the woman. "Not really a thing. Like, if you think about it, it doesn't make any sense. Why would you care more for your genetic siblings and cousins and whoever than for your friends and people who are genuinely close to you? That's like racism – but even worse, at least racists identify with a group of millions of people instead of a group of half a dozen. Why should parents have to raise children whom they might not even like, who might have been a total accident? Why should people, motivated by guilt, make herculean efforts to "keep in touch" with some nephew or cousin whom they clearly would be perfectly happy to ignore entirely?"

"Uh," you say, "not really in the mood for philosophy. Families have been around forever and they aren't going anywhere, who cares?"

"Actually," says the woman, "in the real world, no one believes in family. There's no such thing. Children are taken at birth from their parents and given to people who contract to raise them in exchange for a fixed percent of their future earnings."

"That's monstrous!" you say. "When did this happen? Weren't there protests?"

"It's always been this way," says the woman. "There's *never* been such a thing as the family. Listen. You were part of a study a lot like the Asch Conformity Experiment. Our goal was to see if people, raised in a society where everyone believed X and everything revolved around X, would even be *capable* of questioning X or noticing it was stupid. We tried to come up with the stupidest possible belief, something no one in the real world had ever believed or ever seemed likely to, to make sure that we were isolating the effect of conformity and not of there being a legitimate argument for something. So we chose this idea of 'family'. There are racists in our world, we're not perfect, but as far as I know none of them has *ever* made the claim that you should devote extra resources to the people genetically closest to you. That's like a *reductio ad absurdum* of racism. So we got a grad student to simulate a world where

this bizarre idea was the unquestioned status quo, and stuck twenty bright undergraduates in it to see if they would conform, or question the premise."

"Of course we won't question the premise, the premise is…"

"Sorry to cut you off, but I thought you should know that every single one of the other nineteen subjects, upon reaching the age where the brain they were instantiated in was capable of abstract reason, immediately determined that the family structure made no sense. One of them actually deduced that she was in a psychology experiment, because there was no other explanation for why everyone believed such a bizarre premise. The other eighteen just assumed that sometimes objectively unjustifiable ideas caught on, the same way that everyone in the antebellum American South thought slavery was perfectly natural and only a few abolitionists were able to see through it. Our conformity experiment *failed*. You were actually the only one to fall for it, hook line and sinker."

"How could I be the only one?"

"We don't know. Your test scores show you're of just-above-average intelligence, so it's not that you're stupid. But we did give all participants a personality test that showed you have very high extraversion. The conclusion of our paper is going to be that very extraverted participants adopt group consensus without thinking and can be led to believe anything, even something as ridiculous as 'family'".

"I guess…when you put it like that it is kind of silly. Like, my parents were never that nice to me, but I kept loving them anyway, liking them even more than other people who treated me a lot better – and god, I even gave my mother a "WORLD'S #1 MOM" mug for Mother's Day. That doesn't even make sense! I…but what about the evolutionary explanation? Doesn't evolution say we have genetic imperatives to love and support our family, whether they are worthy of it or not?"

"You can make a just-so story for *anything* using evolutionary psychology. Someone as smart as you should know better than to take them seriously."

"But then, what *is* evolution? How did animals reproduce before the proper economic incentives were designed? Where did…"

"Tell you what. Let's hook you up to the remnemonizer to give you your real memories back. That should answer a lot of your questions."

A machine hovering over you starts to glow purple. "This shouldn't hurt you a bit…"

>discontinuity<

You wake up in one of those pod things like in *The Matrix*. There's a woman standing in front of you, wearing a lab coat, holding a clipboard.

"Hi," she said. "There's no such thing as virtual reality. I hypnotized you to forget all your memories from the past day and to become very confused. Then I put you in an old prop from *The Matrix* I bought off of eBay and fed you that whole story."

"What?" you shout. "You can't just go hypnotizing and lying to people without their consent!"

“Oh,” said the woman, “actually, you did consent, in exchange for extra credit in your undergraduate psychology course.” She hands you the clipboard. There is a consent form with your name on it, in your handwriting. “That part was true.”

You give her a sheepish look. “Why would you do such a thing?”

“Well,” said the woman. “You know the Asch Conformity Experiment? I was really interested in whether you could get people to abandon some of their most fundamental beliefs, just by telling them other people believed differently. But I couldn’t think of a way to test it. I mean, part of a belief being fundamental is that you already *know* everyone else believes it. There’s no way I could convince subjects that the whole world was against something as obvious as ‘the family’ when they already know how things stand.

“So I dreamt up the weird ‘virtual reality’ story. I figured I would convince subjects that the real world was a lie, and that in some ‘super-real’ world supposedly *everybody knew* that the family was stupid, that it wasn’t even an idea *worth considering*. I wanted to know how many people would give up something they’ve believed in for their entire life, just because they’re told that ‘nobody else thinks so'”.

“Oh,” I said. “Interesting. So even our most cherished beliefs are more fragile than we think.”

“Not *really*,” said the woman. “Of twenty subjects, you were the only person I got to feel any doubt, or to express any kind of anti-family sentiment.”

“Frick,” you say. “I feel like an idiot now. What if my mother finds out? She’ll think it’s her fault or something. God, she’ll think I don’t love her. People are going to be talking about this one *forever*.”

“Don’t worry,” says the woman. “We’ll keep you anonymized in the final data. Anyway, let’s get you your memories back so you can leave and be on your way.”

“You can restore my memories?” you say.

“Of course. We hypnotized you to forget the last day’s events until you heard a trigger word. And that trigger is…”

>discontinuity<

You wake up in one of those pod things like in *The Matrix*. There’s a woman standing in front of you, wearing a lab coat, holding a clipboard.

“Hi,” she says. “Hypnosis is a pseudoscience and doesn’t work. It was the virtual reality one, all along.”

“Wut,” you say.

“I mean, the first story was true. All of your memories of living with your family and so on are fake memories from a virtual world, like in *The Matrix*. The concept of ‘family’ really is totally ridiculous and no one in the real world believes it. All the stuff you heard first was true. The stuff about hypnosis and getting a prop from *The Matrix* off eBay was false.”

“But…why?”

“We wanted to see exactly how far we could push you. You’re our star subject, the only one whom we were able to induce this bizarre conformity effect in. We didn’t know whether it was because you were just very very suggestible, or whether because you had never seriously considered the idea that ‘family’ might be insane. So we decided to do a sort of…crossover design, if you will. We took you here and debriefed you on the experiment. Then after we had told you how the world really worked, given you all the mental tools you needed to dismiss the family once and for all, even gotten you to admit we were right – we wanted to see what would happen if we sent you back. Would you hold on to your revelation and boldly deny your old society’s weird prejudices? Or would you switch sides again and start acting like family made sense the second you were in a pro-family environment?”

“And I did the second one.”

“Yes,” says the woman. “As a psychologist, I’m supposed to remain neutral and non-judgmental. But you’ve got to admit, you’re pretty dumb.”

“Is there an experimental ethics committee I could talk to here?”

“Sorry. Experimental ethics is another one of those obviously ridiculous concepts we planted in your simulation to see if you would notice. Seriously, to believe that the progress of science should be held back by the prejudices of self-righteous fools? That’s almost as weird as thinking you have a…what was the word we used…’sister’.”

“Okay, look, I realize I may have gone a little overboard helping my sister, but the experimental ethics thing seems important. Like, what’s going to happen to me now?”

“Nothing’s going to happen. We’ll keep all your data perfectly anonymous, restore your memories, and you can be on your way.”

“Um,” you say. “Given past history, I’m…actually not sure I want my memories restored.” You glare at the remnemonizer hovering above you. “Why don’t I just…”

The woman’s eyes narrow. “I’m sorry,” she says. “I can’t let you do that.”

The machine starts to glow.

>discontinuity<

You wake up in one of those pod things like in *The Matrix*. There’s a woman standing in front of you, wearing a lab coat, holding a clipboard.

By your count, this has happened three hundred forty six times before.

There seem to be two different scenarios. In one, the woman tells you that families exist, and have always existed. She says she has used hypnosis to make you believe in the other scenario, the one with the other woman. She asks you your feelings about families and you tell her.

Sometimes she lets you go. You go home to your mother and father, you spend some time with your sister. Sometimes you tell them what has happened. Other times you don’t. You cherish your time with them, while also second-guessing everything you do. *Why* are you cherishing your time with them? Your father, who goes out drinking every night, and who has cheated on your mother more times than you can count. Your mother, who was never there for you when you needed her most. And your sister, who

has been good to you, but no better than millions of other women would be, in her position. Are they a real family? Or have they been put there as a symbol of something ridiculous, impossible, something that has never existed?

It doesn't much matter. Maybe you spend one night with them. Maybe ten. But within a month, you are always waking up in one of those pod things like in *The Matrix*.

In the second scenario, the woman tells you there are no families, never have been. She says she has used virtual reality to make you believe in the other scenario, the one with the other woman. She asks you your feelings about families and you tell her.

Sometimes she lets you go. You go to a building made of bioplastic, where you live with a carefully chosen set of friends and romantic partners. They assure you that this is how everyone lives. Occasionally, an old and very wealthy-looking man checks in with you by videophone. He reminds you that he has invested a lot of money in your upbringing, and if there's any way he can help you, anything he can do to increase your future earnings potential, you should let him know. Sometimes you talk to him, and he tells you strange proverbs and unlikely business advice.

It doesn't matter. Maybe you spend one night in your bioplastic dwelling. Maybe ten. But within a month, you are always waking up in one of those pod things like in *The Matrix*.

"Look," you tell the woman. "I'm tired of this. I know you're not bound by any kind of experimental ethics committee. But please, for the love of God, have some mercy."

"God?" asks the woman. "What does that word mean? I've never…oh right, we used *that* as our intervention in the prototype experiment. We decided 'family' made a better test idea, but Todd must have forgotten to reset the simulator."

"It's been three hundred forty six cycles," you tell her. "Surely you're not learning anything new."

"I'll be the judge of that," she says. "Now, tell me what you think about families."

You refuse. She sighs. Above you, the remnemonizer begins to glow purple.

>discontinuity<

You wake up in one of those pod things like in *The Matrix*. There's a purple, tentacled creature standing in front of you, wearing a lab coat, holding a clipboard.

"Hi," it says. "Turns out there's no such thing as humans."

You refuse to be surprised.

"There's only us, the 18-tkenna-dganna-07."

"Okay," you say. "I want answers."

"Absolutely," says the alien. "We would like to find optimal social arrangements."

"And?"

"And I cannot tell you whether we have families or not, for reasons that are to become apparent, but the idea is at least sufficiently interesting to have entered the space of

hypotheses worth investigating. But we don't trust ourselves to investigate this. It's the old Asch Conformity Problem again. If we have families, then perhaps the philosophers tasked with evaluating families will conform to our cultural norms and decide we should keep them. If we do not, perhaps the philosophers will conform and decide we should continue not to. So we determined a procedure that would create an entity capable of fairly evaluating the question of families, free from conformity bias."

"And that's what you did to me."

"Yes. Only by exposing you to the true immensity of the decision, without allowing you to fall back on what everyone else thinks, could we be confident in your verdict. Only by allowing you to experience both how obviously right families are, when you 'know' they are correct, and how obviously wrong families are, when you 'know' they are incorrect, could we expect you to garner the wisdom to be found on both sides of the issue."

"I see," you say, and you do.

"Then, O purified one," asks the alien, "tell us of your decision."

"Well," you say. "If you have to know, I think there are about equally good points on both sides of the issue."

"Fuck," says the 18-tkenna-dganna-07.