AI competitions don’t produce useful models

Originalartikel

Backup

<html> A huge new CT brain dataset was released today, with the goal of training models to detect intracranial haemorrhage. So far, it looks pretty good, although I haven’t dug into it in detail yet (and the devil is often in the detail). The dataset has been released for a competition, which obviously lead to the usual friendly rivalry on Twitter: Of course, this lead to cynicism from the usual suspects as well. <img data-attachment-id=„8689“ data-permalink=„https://lukeoakdenrayner.wordpress.com/2019/09/19/ai-competitions-dont-produce-useful-models/ai-competition2/“ data-orig-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/ai-competition2.png“ data-orig-size=„593,561“ data-comments-opened=„1“ data-image-meta=„{"aperture":"0","credit":"","camera":"","caption":"","created_timestamp":"0","copyright":"","focal_length":"0","iso":"0","shutter_speed":"0","title":"","orientation":"0"}“ data-image-title=„ai competition2“ data-image-description=„“ data-medium-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/ai-competition2.png?w=300“ data-large-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/ai-competition2.png?w=593“ class=„aligncenter wp-image-8689“ src=„https://lukeoakdenrayner.files.wordpress.com/2019/09/ai-competition2.png?w=444&h=420“ alt=„ai competition2“ width=„444“ height=„420“

srcset=„https://lukeoakdenrayner.files.wordpress.com/2019/09/ai-competition2.png?w=444&h=420 444w, https://lukeoakdenrayner.files.wordpress.com/2019/09/ai-competition2.png?w=150&h=142 150w, https://lukeoakdenrayner.files.wordpress.com/2019/09/ai-competition2.png?w=300&h=284 300w, https://lukeoakdenrayner.files.wordpress.com/2019/09/ai-competition2.png 593w“ sizes=„(max-width: 444px) 100vw, 444px“/> And the conversation continued from there, with thoughts ranging from “but since there is a hold out test set, how can you overfit?” to “the proposed solutions are never intended to be applied directly” (the latter from a previous competition winner). As the discussion progressed, I realised that while we “all know” that competition results are more than a bit dubious in a clinical sense, I’ve never really seen a compelling explanation for why this is so. Hopefully that is what this post is, an explanation for why competitions are not really about building useful AI systems. <hr/><h5>DISCLAIMER: I originally wrote this post expecting it to be read by my usual readers, who know my general positions on a range of issues. Instead, it was spread widely on Twitter and HackerNews, and it is pretty clear that I didn’t provide enough context for a number of statements made. I am going to write a follow-up to clarify several things, but as a quick response to several common criticisms:</h5> <h5>I don’t think AlexNet is a better model than ResNet. That position would be ridiculous, particularly given all of my published work uses resnets and densenets, not AlexNets.</h5>   <h5>I think this miscommunication came from me not defining my terms: a “useful” model would be one that works for the task it was trained on. It isn’t a model architecture. If architectures are developed in the course of competitions that are broadly useful, then that is a good architecture, but the particular implementation submitted to the competition is not necessarily a useful model.</h5>   <h5>The stats in this post are wrong, but they are meant to be wrong in the right direction. They are intended for illustration of the concept of crowd-based overfitting, not accuracy. Better approaches would almost all require information that isn’t available in public leaderboards. I may update the stats at some point to make them more accurate, but they will never be perfect.</h5>   <h5>I was trying something new with this post – it was a response to a Twitter conversation, so I wanted to see if I could write it in one day to keep it contemporaneous. Given my usual process is spending several weeks and many rewrites per post, this was a risk. I think the post still serves its purpose, but I don’t personally think the risk paid off. If I had taken even another day or two, I suspect I would have picked up most of these issues before publication. Mea culpa.</h5> <hr/> <h3>Let’s have a battle</h3> <img data-attachment-id=„8691“ data-permalink=„https://lukeoakdenrayner.wordpress.com/2019/09/19/ai-competitions-dont-produce-useful-models/748284__safe_animated_screencap_equestriagirls_rainbowrocks_spoiler-colon-rainbowrocks_sonatadusk_adagiodazzle_ariablaze_thedazzlings/“ data-orig-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/748284__safe_animated_screencap_equestriagirls_rainbowrocks_spoiler-colon-rainbowrocks_sonatadusk_adagiodazzle_ariablaze_thedazzlings.gif“ data-orig-size=„576,324“ data-comments-opened=„1“ data-image-meta=„{"aperture":"0","credit":"","camera":"","caption":"","created_timestamp":"0","copyright":"","focal_length":"0","iso":"0","shutter_speed":"0","title":"","orientation&

quot;:"0"}“ data-image-title=„748284safe_animated_screencap_equestriagirls_rainbowrocks_spoiler-colon-rainbowrocks_sonatadusk_adagiodazzle_ariablaze_thedazzlings“ data-image-description=„“ data-medium-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/748284__safe_animated_screencap_equestriagirls_rainbowrocks_spoiler-colon-rainbowrocks_sonatadusk_adagiodazzle_ariablaze_thedazzlings.gif?w=300“ data-large-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/748284__safe_animated_screencap_equestriagirls_rainbowrocks_spoiler-colon-rainbowrocks_sonatadusk_adagiodazzle_ariablaze_thedazzlings.gif?w=576“ class=„aligncenter size-full wp-image-8691“ src=„https://lukeoakdenrayner.files.wordpress.com/2019/09/748284__safe_animated_screencap_equestriagirls_rainbowrocks_spoiler-colon-rainbowrocks_sonatadusk_adagiodazzle_ariablaze_thedazzlings.gif?w=656“ alt=„748284safe_animated_screencap_equestriagirls_rainbowrocks_spoiler-colon-rainbowrocks_sonatadusk_adagiodazzle_ariablaze_thedazzlings“/> <h5 class=„c1“>Nothing wrong with a little competition.*</h5> So what is a competition in medical AI? Here are a few options: getting teams to try to solve a clinical problem getting teams to explore how problems might be solved and to try novel solutions getting teams to build a model that performs the best on the competition test set a waste of time Now, I’m not so jaded that I jump to the last option (what is valuable to spend time on is a matter of opinion, and clinical utility is only one consideration. More on this at the end of the article). But what about the first three options? Do these models work for the clinical task, and do they lead to broadly applicable solutions and novelty, or are they only good in the competition and not in the real world? (Spoiler: I’m going to argue the latter). <hr/> <h3>Good models and bad models</h3> Should we expect this competition to produce good models? Let’s see what one of the organisers says. Cool. Totally agree. The lack of large, well-

labeled datasets is the biggest major barrier to building useful clinical AI, so this dataset should help. But saying that the dataset can be useful is not the same thing as saying the competition will produce good models. So to define our terms, let’s say that a good model is a model that can detect brain haemorrhages on unseen data (cases that the model has no knowledge of). So conversely, a bad model is one that doesn’t detect brain haemorrhages in unseen data. These definitions will be non-controversial. Machine Learning 101. I’m sure the contest organisers agree with these definitions, and would prefer their participants to be producing good models rather than bad models. In fact, they have clearly set up the competition in a way designed to promote good models. It just isn’t enough. <hr/> <h3>Epi vs ML, FIGHT!</h3> <img data-attachment-id=„8715“
data-
permalink=„https://lukeoakdenrayner.wordpress.com/2019/09/19/ai-competitions-dont-produce-useful-models/source-2/“ data-orig-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/source.gif“ data-orig-size=„400,225“ data-comments-opened=„1“ data-image-meta=„{"aperture":"0","credit":"","camera":"","caption":"","created_timestamp":"0","copyright":"","focal_length":"0","iso":"0","shutter_speed":"0","title":"","orientation":"0"}“ data-image-title=„source“ data-image-description=„“ data-medium-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/source.gif?w=300“ data-large-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/source.gif?w=400“ class=„alignnone size-full wp-image-8715“
src=„https://lukeoakdenrayner.files.wordpress.com/2019/09/source.gif?w=656“ alt=„source“/> <h5 class=„c1“>If only academic arguments were this cute</h5> ML101 (now personified) tells us that the way to control overfitting is to use a hold-out test set, which is data that has not been seen during model training. This simulates seeing new patients in a clinical setting. ML101 also says that hold-out data is only good for one test. If you test multiple models, then even if you don’t cheat and leak test information into your development process, your best result is probably an outlier which was only better than your worst result by chance. So competition organisers these days produce hold-out test sets, and only let each team run their model on the data once. Problem solved, says ML101. The winner only tested once, so there is no reason to think they are an outlier, they just have the best model. Not so fast, buddy. Let me introduce you to Epidemiology 101, who claims to have a magic coin. <img data-attachment-id=„8704“
data-
permalink=„https://lukeoakdenrayner.wordpress.com/2019/09/19/ai-competitions-dont-produce-useful-models/fsnuw6sujx28bh5itvip/“
data-orig-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/fsnuw6sujx28bh5itvip.gif“ data-orig-size=„636,287“ data-comments-opened=„1“ data-image-meta=„{"aperture":"0","credit":"","camera":"","caption":"","created_timestamp":"0","copyright":"","focal_length":"0","iso":"0","shutter_speed":"0","title":"","orientation":"0"}“ data-image-title=„fsnuw6sujx28bh5itvip“ data-image-description=„“ data-medium-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/fsnuw6sujx28bh5itvip.gif?w=300“
data-

large-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/fsnuw6sujx28bh5itvip.gif?w=636“ class=„alignnone size-medium wp-image-8704“ src=„https://lukeoakdenrayner.files.wordpress.com/2019/09/fsnuw6sujx28bh5itvip.gif?w=300&h=135“ alt=„fsnuw6sujx28bh5itvip“ width=„300“ height=„135“/> Epi101 tells you to flip the coin 10 times. If you get 8 or more heads, that confirms the coin is magic (while the assertion is clearly nonsense, you play along since you know that 8/10 heads equates to a p-value of <0.05 for a fair coin, so it must be legit). Unbeknownst to you, Epi101 does the same thing with 99 other people, all of whom think they are the only one testing the coin. What do you expect to happen? If the coin is totally normal and not magic, around 5 people will find that the coin is special. Seems obvious, but think about this in the context of the individuals. Those 5 people all only ran a single test. According to them, they have statistically significant evidence they are holding a “magic” coin. Now imagine you aren’t flipping coins. Imagine you are all running a model on a competition test set. Instead of wondering if your coin is magic, you instead are hoping that your model is the best one, about to earn you $25,000. Of course, you can’t submit more than one model. That would be cheating. One of the models could perform well, the equivalent of getting 8 heads with a fair coin, just by chance. Good thing there is a rule against it submitting multiple models, or any one of the other 99 participants and their 99 models could win, just by being lucky… <img data-attachment-id=„8711“ data-permalink=„https://lukeoakdenrayner.wordpress.com/2019/09/19/ai-competitions-dont-produce-useful-models/tumblr_mvn6uwbyep1shi48to5_250/“ data-orig-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/tumblr_mvn6uwbyep1shi48to5_250.gif“ data-orig-size=„245,150“ data-comments-opened=„1“ data-image-meta=„{"aperture":"0","credit":"","camera":"","caption":"","created_timestamp":"0","copyright":"","focal_length":"0","iso":"0

","shutter_speed":"0","title":"","orientation":"0"}“ data-image-title=„tumblr_mvn6uwbyep1shi48to5_250“ data-image-description=„“
data-medium-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/tumblr_mvn6uwbyep1shi48to5_250.gif?w=245“
data-large-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/tumblr_mvn6uwbyep1shi48to5_250.gif?w=245“ class=„alignnone size-full wp-image-8711“
src=„https://lukeoakdenrayner.files.wordpress.com/2019/09/tumblr_mvn6uwbyep1shi48to5_250.gif?w=656“ alt=„tumblr_mvn6uwbyep1shi48to5_250“/> <hr/> <h3>Multiple hypothesis testing</h3> The effect we saw with Epi101’s coin applies to our competition, of course. Due to random chance, some percentage of models will outperform other ones, even if they are all just as good as each other. Maths doesn’t care if it was one team that tested 100 models, or 100 teams. Even if certain models are better than others in a meaningful sense^, unless you truly believe that the winner is uniquely able to ML-wizard, you have to accept that at least some other participants would have achieved similar results, and thus the winner only won because they got lucky. The real “best performance” will be somewhere back in the pack, probably above average but below the winner^^. <img data-attachment-id=„8709“
data-permalink=„https://lukeoakdenrayner.wordpress.com/2019/09/19/ai-competitions-dont-produce-useful-models/ai-competition4-2/“
data-orig-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/ai-competition4-1.png“ data-orig-size=„680,341“ data-comments-opened=„1“ data-image-meta=„{"aperture":"0","credit":"","camera":"","caption":"","created_timestamp":"0","copyright":"","focal_length":"0","iso":"0","shutter_speed":"0","title":"","orientation":"0"}“ data-image-title=„ai competition4“ data-image-description=„“ data-medium-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/ai-competition4-1.png?w=300“ data-large-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/ai-competition4-1.png?w=656“ class=„wp-image-8709 aligncenter“
src=„https://lukeoakdenrayner.files.wordpress.com/2019/09/ai-competition4-1.png?w=620&h=311“ alt=„ai competition4“ width=„620“ height=„311“
srcset=„https://lukeoakdenrayner.files.wordpress.com/2019/09/ai-competition4-1.png?w=620&h=311 620w,
https://lukeoakdenrayner.files.wordpress.com/2019/09/ai-competition4-1.png?w=150&h=75 150w,
https://lukeoakdenrayner.files.wordpress.com/2019/09/ai-competition4-1.png?w=300&h=150 300w, https://lukeoakdenrayner.files.wordpress.com/2019/09/ai-competition4-1.png 680w“
sizes=„(max-width: 620px) 100vw, 620px“/> Epi101 says this effect is called multiple hypothesis testing. In the case of a competition, you have a ton of hypotheses – that each participant was better than all others. For 100 participants, 100 hypotheses. One of those hypotheses, taken in isolation, might show us there is a winner with statistical significance (p<0.05). But taken together, even if the winner has a calculated “winning” p-value of less than 0.05, that doesn’t mean we only have a 5% chance of making an unjustified decision. In fact, if this was coin flips (which is easier to calculate but not absurdly different), we would

have a greater than 99% chance that one or more people would “win” and come up with 8 heads! That is what an AI competition winner is; an individual who happens to get 8 heads while flipping fair coins. Interestingly, while ML101 is very clear that running 100 models yourself and picking the best one will result in overfitting, they rarely discuss this “overfitting of the crowds”. Strange, when you consider that almost all ML research is done of heavily over-tested public datasets … So how do we deal with multiple hypothesis testing? It all comes down to the cause of the problem, which is the data. Epi101 tells us that any test set is a biased version of the target population. In this case, the target population is “all patients with CT head imaging, with and without intracranial haemorrhage”. Let’s look at how this kind of bias might play out, with a toy example of a small hypothetical population: <img data-attachment-id=„8716“
data-
permalink=„https://lukeoakdenrayner.wordpress.com/2019/09/19/ai-competitions-dont-produce-useful-models/brain-bleeds-1/“
data-orig-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/brain-bleeds-1.png“ data-orig-size=„1341,277“ data-comments-opened=„1“ data-image-
meta=„{"aperture":"0","credit":"","camera":"","caption":"","created_timestamp":"0","copyright":"","focal_length":"0","iso":"0","shutter_speed":"0","title":"","orientation":"0"}“ data-image-title=„brain bleeds 1“ data-image-description=„“ data-medium-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/brain-bleeds-1.png?w=300“ data-large-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/brain-bleeds-1.png?w=656“ class=„wp-image-8716 aligncenter“
src=„https://lukeoakdenrayner.files.wordpress.com/2019/09/brain-bleeds-1.png?w=675&h=140“
alt=„brain bleeds 1“ width=„675“ height=„140“
srcset=„https://lukeoakdenrayner.files.wordpress.com/2019/09/brain-bleeds-1.png?w=675&h=140 675w,
https://lukeoakdenrayner.files.wordpress.com/2019/09/brain-bleeds-1.png?w=150&h=31 150w,
https://lukeoakdenrayner.files.wordpress.com/2019/09/brain-bleeds-1.png?w=300&h=62 300w,
https://lukeoakdenrayner.files.wordpress.com/2019/09/brain-bleeds-1.png?w=768&h=159 768w,
https://lukeoakdenrayner.files.wordpress.com/2019/09/brain-bleeds-1.png?w=1024&h=212 1024w, https://lukeoakdenrayner.files.wordpress.com/2019/09/brain-bleeds-1.png 1341w“
sizes=„(max-width: 675px) 100vw, 675px“/> In this population, we have a pretty reasonable “clinical” mix of cases. 3 intra-cerebral bleeds (likely related to high blood pressure or stroke), and two traumatic bleeds (a subdural on the right, and an extradural second from the left). Now let’s sample this population to build our test set: <img data-attachment-id=„8717“
data-
permalink=„https://lukeoakdenrayner.wordpress.com/2019/09/19/ai-competitions-dont-produce-useful-models/brain-bleeds-2/“
data-orig-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/brain-bleeds-2.png“ data-orig-size=„782,272“ data-comments-opened=„1“ data-image-
meta=„{"aperture":"0","credit":"","camera":"","caption":"","created_timestamp":"0","copyright":"","focal_length":"0","iso":"0","shutter_speed":"0","title":"","orientation":"0"}“ data-image-title=„brain bleeds 2“ data-image-description=„“ data-medium-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/brain-bleeds-2.png?w=300“ data-large-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/brain-bleeds-2.png?w=656“ class=„wp-image-8717 aligncenter“

src=„https://lukeoakdenrayner.files.wordpress.com/2019/09/brain-bleeds-2.png?w=388&h=135“ alt=„brain bleeds 2“ width=„388“ height=„135“ srcset=„https://lukeoakdenrayner.files.wordpress.com/2019/09/brain-bleeds-2.png?w=388&h=135 388w, https://lukeoakdenrayner.files.wordpress.com/2019/09/brain-bleeds-2.png?w=776&h=270 776w, https://lukeoakdenrayner.files.wordpress.com/2019/09/brain-bleeds-2.png?w=150&h=52 150w, https://lukeoakdenrayner.files.wordpress.com/2019/09/brain-bleeds-2.png?w=300&h=104 300w, https://lukeoakdenrayner.files.wordpress.com/2019/09/brain-bleeds-2.png?w=768&h=267 768w“ sizes=„(max-width: 388px) 100vw, 388px“/> Randomly, we end up with mostly extra-axial (outside of the brain itself) bleeds. A model that performs well on this test will not necessarily work as well on real patients. In fact, you might expect a model that is really good at extra-axial bleeds at the expense of intra-cerebral bleeds to win. But Epi101 doesn’t only point out problems. Epi101 has a solution. <hr/> <h3>So powerful</h3> There is only one way to have an unbiased test set – if it includes the entire population! Then whatever model does well in the test will also be the best in practice, because you tested it on all possible future patients (which seems difficult). This leads to a very simple idea – your test results become more reliable as the test set gets larger. We can actually predict how reliable test sets are using power calculations. These are power curves. If you have a rough idea of how much better your “winning” model will be than the next best model, you can estimate how many test cases you need to reliably show that it is better. So to find out if you model is 10% better than a competitor, you would need about 300 test cases. You can also see how exponentially the number of cases needed grows as the difference between models gets narrower. Let’s put this into practice. If we look at another medical AI competition, the SIIM-ACR pneumothorax segmentation challenge, we see that the difference in Dice scores (ranging between 0 and 1) is negligible at the top of the leaderboard. Keep in mind that this competition had a dataset of 3200 cases (and that is being generous, they don’t all contribute to the Dice score

equally). <img data-attachment-id=„8719“
data-
permalink=„https://lukeoakdenrayner.wordpress.com/2019/09/19/ai-competitions-dont-produce-useful-models/pneumo-challenge/“
data-orig-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/pneumo-challenge.png“ data-orig-size=„1209,949“ data-comments-opened=„1“ data-image-meta=„{"aperture":"0","credit":"","camera":"","caption":"","created_timestamp":"0","copyright":"","focal_length":"0","iso":"0","shutter_speed":"0","title":"","orientation":"0"}“ data-image-title=„pneumo challenge“ data-image-description=„“ data-medium-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/pneumo-challenge.png?w=300“
data-
large-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/pneumo-challenge.png?w=656“
class=„wp-image-8719 aligncenter“
src=„https://lukeoakdenrayner.files.wordpress.com/2019/09/pneumo-challenge.png?w=663&h=521“ alt=„pneumo challenge“ width=„663“ height=„521“
srcset=„https://lukeoakdenrayner.files.wordpress.com/2019/09/pneumo-challenge.png?w=663&h=521 663w,
https://lukeoakdenrayner.files.wordpress.com/2019/09/pneumo-challenge.png?w=150&h=118 150w,
https://lukeoakdenrayner.files.wordpress.com/2019/09/pneumo-challenge.png?w=300&h=235 300w,
https://lukeoakdenrayner.files.wordpress.com/2019/09/pneumo-challenge.png?w=768&h=603 768w,
https://lukeoakdenrayner.files.wordpress.com/2019/09/pneumo-challenge.png?w=1024&h=804 1024w, https://lukeoakdenrayner.files.wordpress.com/2019/09/pneumo-challenge.png 1209w“
sizes=„(max-width: 663px) 100vw, 663px“/> So the difference between the top two was 0.0014 … let’s chuck that into a sample size calculator. <img data-attachment-id=„8721“
data-
permalink=„https://lukeoakdenrayner.wordpress.com/2019/09/19/ai-competitions-dont-produce-useful-models/sample-size-1-2/“
data-orig-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/sample-size-1-1.png“ data-orig-size=„789,523“ data-comments-opened=„1“ data-image-meta=„{"aperture":"0","credit":"","camera":"","caption":"","created_timestamp":"0","copyright":"","focal_length":"0","iso":"0","shutter_speed":"0","title":"","orientation":"0"}“ data-image-title=„sample size 1“ data-image-description=„“ data-medium-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/sample-size-1-1.png?w=300“ data-large-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/sample-size-1-1.png?w=656“ class=„wp-image-8721 aligncenter“
src=„https://lukeoakdenrayner.files.wordpress.com/2019/09/sample-size-1-1.png?w=498&h=330“ alt=„sample size 1“ width=„498“ height=„330“
srcset=„https://lukeoakdenrayner.files.wordpress.com/2019/09/sample-size-1-1.png?w=498&h=330 498w,
https://lukeoakdenrayner.files.wordpress.com/2019/09/sample-size-1-1.png?w=150&h=99 150w,
https://lukeoakdenrayner.files.wordpress.com/2019/09/sample-size-1-1.png?w=300&h=199 300w,
https://lukeoakdenrayner.files.wordpress.com/2019/09/sample-size-1-1.png?w=768&h=509

768w, https://lukeoakdenrayner.files.wordpress.com/2019/09/sample-size-1-1.png 789w“ sizes=„(max-width: 498px) 100vw, 498px“/> Ok, so to show a significant difference between these two results, you would need 920,000 cases. But why stop there? We haven’t even discussed multiple hypothesis testing yet. This absurd number of cases needed is simply if there was ever only one hypothesis, meaning only two participants. If we look at the leaderboard, there were 351 teams who made submissions. The rules say they could submit two models, so we might as well assume there were at least 500 tests. This has to produce some outliers, just like 500 people flipping a fair coin. Epi101 to the rescue. Multiple hypothesis testing is really common in medicine, particularly in “big data” fields like genomics. We have spent the last few decades learning how to deal with this. The simplest reliable way to manage this problem is called the Bonferroni correction^^. The Bonferroni correction is super simple: you divide the p-value by the number of tests to find a “statistical significance threshold” that has been adjusted for all those extra coin flips. So in this case, we do 0.05/500. Our new p-value target is 0.0001, any result worse than this will be considered to support the null hypothesis (that the competitors performed equally well on the test set). So let’s plug that in our power calculator. Cool! It only increased a bit… to 2.6 million cases needed for a valid result :p Now, you might say I am being very unfair here, and that there must be some small group of good models at the top of the leaderboard that are not clearly different from each other^^^. Fine, lets be generous. Surely no-one will complain if I compare the 1st place model to the 150th model? <img data-attachment-id=„8726“ data-permalink=„https://lukeoakdenrayner.wordpress.com/2019/09/19/ai-competitions-dont-produce-useful-models/sample-size-3/“ data-orig-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/sample-size-3-1.png“ data-orig-size=„756,500“ data-comments-opened=„1“ data-image-meta=„{"aperture":"0","credit":"","camera":"","caption":"","created_timestamp":"0",&q

uot;copyright":"","focal_length":"0","iso":"0","shutter_speed":"0","title":"","orientation":"0"}“ data-image-title=„sample size 3“ data-image-description=„“ data-medium-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/sample-size-3-1.png?w=300“ data-large-file=„https://lukeoakdenrayner.files.wordpress.com/2019/09/sample-size-3-1.png?w=656“ class=„wp-image-8726 aligncenter“ src=„https://lukeoakdenrayner.files.wordpress.com/2019/09/sample-size-3-1.png?w=480&h=317“ alt=„sample size 3“ width=„480“ height=„317“ srcset=„https://lukeoakdenrayner.files.wordpress.com/2019/09/sample-size-3-1.png?w=480&h=317 480w, https://lukeoakdenrayner.files.wordpress.com/2019/09/sample-size-3-1.png?w=150&h=99 150w, https://lukeoakdenrayner.files.wordpress.com/2019/09/sample-size-3-1.png?w=300&h=198 300w, https://lukeoakdenrayner.files.wordpress.com/2019/09/sample-size-3-1.png 756w“ sizes=„(max-width: 480px) 100vw, 480px“/>

So still more data than we had. In fact, I have to go down to the 192nd placeholder to find a result where the sample size was enough to produce a “statistically significant” difference. But maybe this is specific to the pneumothorax challenge? What about other competitions? In MURA, we have a test set of 207 x-rays, with 70 teams submitting “no more than two models per month”, so lets be generous and say 100 models were submitted. Running the numbers, the “first place” model is only significant versus the 56th placeholder and below. In the RSNA Pneumonia Detection Challenge, there were 3000 test images with 350 teams submitting one model each. The first place was only significant compared to the 30th place and below. And to really put the cat amongst the pigeons, what about outside of medicine? As we go left to right in ImageNet results, the improvement year on year slows (the effect size decreases) and the number of people who have tested on the dataset increases. I can’t really estimate the numbers, but knowing what we know about multiple testing does anyone really believe the SOTA rush in the mid 2010s was anything but

crowdsourced overfitting? <hr/> <h3>So what are competitions for?</h3> They obviously aren’t to reliably find the best model. They don’t even really reveal useful techniques to build great models, because we don’t know which of the hundred plus models actually used a good, reliable method, and which method just happened to fit the under-powered test set. You talk to competition organisers … and they mostly say that competitions are for publicity. And that is enough, I guess. AI competitions are fun, community building, talent scouting, brand promoting, and attention grabbing. But AI competitions are not to develop useful models.

<hr/><h5>* I have a young daughter, don’t judge me for my encyclopaedic knowledge of My Little Pony.**</h5> <h5>** not that there is anything wrong with My Little Pony**.**
Friendship is magic. There is just an unsavoury internet element that matches my demographic who is really into the show. I’m no brony.</h5> <h5>*</h5> barring the near complete white-washing of a children’s show about multi-coloured horses.</h5> <h5>^ we can actually understand model performance with our coin analogy. Improving the model would be equivalent to bending the coin. If you are good at coin bending, doing this will make it more likely to land on heads, but unless it is 100% likely you still have no guarantee to “win”. If you have a 60%-chance-of-heads coin, and everyone else has a 50% coin, you objectively have the best coin, but your chance of getting 8 heads out of 10 flips is still only 17%. Better than the 5% the rest of the field have, but remember that there are 99 of them. They have a cumulative chance of over 99% that one of them will get 8 or more heads.</h5> <h5>^^ people often say the Bonferroni correction is a bit conservative, but remember, we are coming in skeptical that these models are actually different from each other. We should be conservative.</h5> <h5>^^^ do please note, the top model here got $30,000 and the second model got nothing. The competition organisers felt that the distinction was reasonable.</h5> </html>