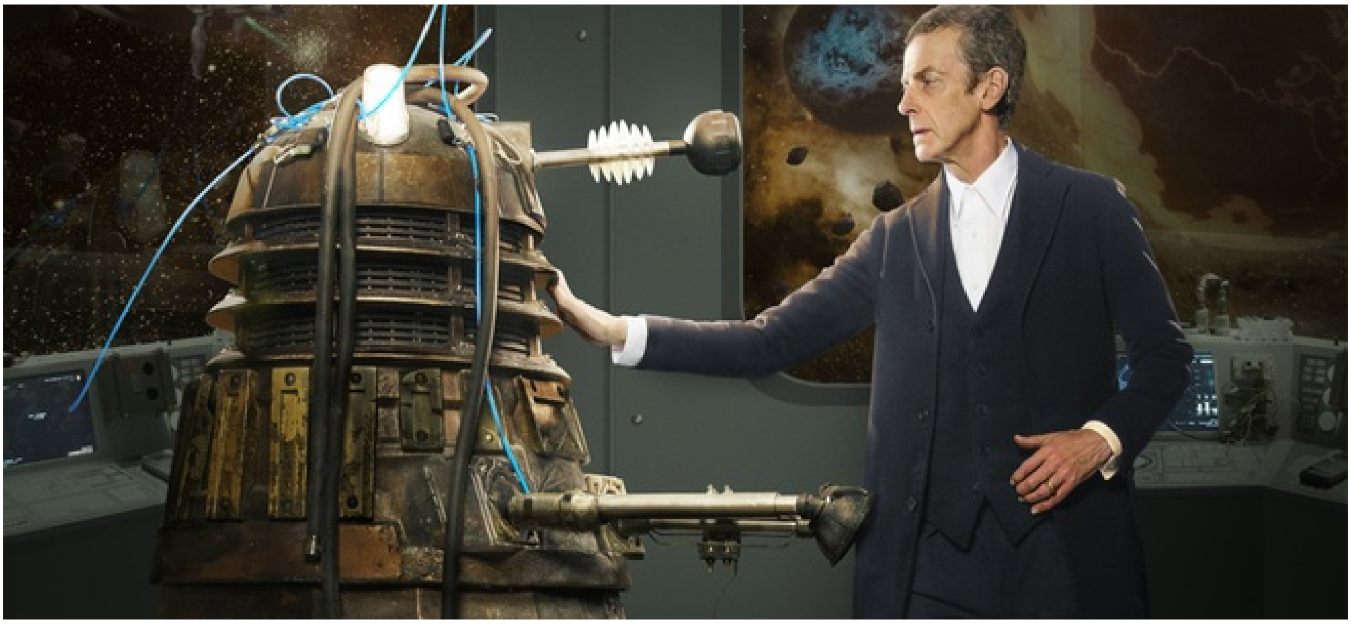
Do machines actually beat doctors?

lukeoakdenrayner.wordpress.com /2016/11/27/do-computers-already-outperform-doctors/

Luke Oakden-Rayner

27/11/2016



If you ask academic machine learning experts about the things that annoy them, high up the list is going to be overblown headlines about how machines are beating humans at some task *where that is completely untrue*. This is partially because reality is already so damn amazing there is no need for hyperbole. AlphaGo beat Lee Sedol convincingly. Most of Atari is solved. Professional transcriptionists lose to voice recongition systems.

Object recognition has been counted on the machine side of the tally for years (albeit with a few more reservations).

But not medicine. Not yet.

Considering the headlines we see, this may surprise many people. For someone who watches the medical AI space, it seems like a day can't go by without some new article reporting on a new piece of research in which the journalists say machines are outperforming human doctors. I'm sure anyone who stumbles on this blog has seen many of them.

A few examples:

Computer Program Beats Doctors at Distinguishing Brain Tumours from Radiation Changes (Neuroscience

News, 2016)

Computers trounce pathologists in predicting lung cancer type, severity (Stanford Medicine News Centre, 2016)

Artificial Intelligence Reads Mammograms With 99% Accuracy (Futurism, 2016)

Digital Diagnosis: Intelligent Machines Do a Better Job Than Humans (Singularity Hub, 2016)

I didn't even have to search for these. Almost all of them are still at the top of my Twitter feed.

Now, these are pretty compelling headlines. The second one is even from the Stanford Uni press, not a clickbait farm. But I hope I can explain why they are both reasonably true statements, but also completely wrong. I think this could be useful for a lot of people, and not just layfolk. Courtesy of Reddit, apparently even some researchers in the field think the machines are already winning.

So I am writing this survival guide: **How to read the medical AI reports with a critical eye, and see the truth through the hype.**

Because the truth is already amazing and beautiful. We don't need the varnish.

The three traps of medical AI articles

There are three major ways these articles get it wrong. They either don't understand medicine, they don't understand AI, or they don't actually compare doctors and machines.

---

1) Humans don't do that

The first one is the most important, because this afflicts healthcare technologists as much as journalists. It is also the most common, and therefore the major culprit behind these headlines.

Journalists, technologists, futurists and so on mostly **don't understand medicine**.

Medicine is **complex**. The biology, the therapeutics, the whole system is so vast it is beyond the scope of any one human mind. Doctors and other healthcare professionals get a feel for it, in a vague and nebulous way, but even that is emphemeral. Some tidbits to remind us how complex treating people actually is:

- We train for 12 years *minimum* to become specialists in a *subfield* of medicine. Doctors are required by law to keep learning throughout their careers, and only hit peak performance after decades.
- Researchers dedicate their lives to tiny fractions of human biology.
- For every doctor or manager there are thousands of other highly trained personnel keeping everything going. In many countries healthcare employs more people than any other industry, and most of them have been through tertiary education.

Medicine is **massive**. Medical research output is larger than any other discipline by orders of magnitude. The scale is mindboggling.

- You think NIPS is getting cramped with a few thousand visitors? The biggest conference in radiology, RSNA (on this weekend), has *over fifty thousand* attendees.
- The impact factor of our top journal is nearly sixty. It has over six hundred thousand readers. The Proceedings of NIPS is under 5. A few teams publish in Nature, sure, but even Nature is only 38.
- Funding totals are hard to pin down, but public funding in the US is at a ratio of something like 3 to 1. For medicine and *all of the rest of science*.
- In PubMed alone (which only indexes the top 4000 or so journals) there are something like a million

medical articles indexed per year.

Medicine is **idiosyncratic.** Most of it has grown around a questionable evidence base. Wrong results, misinterpreted results, unreproducible results, no results. Many of our decisions are made for non-scientific reasons, guided by culture, politics, finance, the law. Unless you have been inside it, it is unlikely you will understand it very well. Even from the inside it doesn't make much sense.

This isn't bragging. Medicine is a mess.

This is explaining why your intuitions about what doctors do are mostly wrong. Just because something *sounds* medical, and seems to be in the scope of medical practice, it doesn't mean doctors are actually doing it.

And if doctors don't do it, learn it, get good at it, *value* it … is it useful to say that machines are better at it?

Let's have a look at some examples.

The article I was recommended in the reddit thread is a pretty famous one. It is a great article from a great research team, and a very valuable contribution that has been taken massively out of context.

It was publicised as Computers trounce pathologists in predicting lung cancer type, severity by the Stanford Medicine News Centre. Fighting talk for sure! Apparently the machine learning system they created has vastly outstripped human pathologists in some sort of predictive task. I am going to ignore the multiple medical errors in the piece, and focus on the meat of the problem. They say computers are better at **predicting** something about cancer.

This is where the alarm bells should ring. If you see the word "predictive" in the headline, you can almost stop there.

Rule 1: Doctors don't do prediction.

This is completely unintuitive, but almost always true. Let's use this article as an example.

> *"Pathology as it is practiced now is very subjective," said Michael Snyder, PhD, professor and chair of genetics. "Two highly skilled pathologists assessing the same slide will agree only about 60 percent of the time."*

So, skilled doctors are pretty much tossing coins in this task. That doesn't sound right. Maybe we should check the research article itself, maybe the journalists just got it wrong? Here is line two of the abstract.

> *However, human evaluation of pathology slides **cannot accurately** predict patients' prognoses.*

Emphasis mine.

Is it clear yet?

Humans *don't do this task*.

They trained a computer to identify which patients with cancer will survive for shorter times. That sounds medical, right? It sounds useful, right?

It isn't. We have no evidence it can help.

Pathologists have trained to provide an answer that will alter treatment choices. Surgery or no surgery. Chemotherapy or radiotherapy. Both, all three, none. These are not the same thing as defining how long someone will live, and there has been no reason to get good at the latter.

Will doing prognosis research help? For sure. I'm fully on board with the Stanford research team here. This is the future of medicine. Predictive analysis is a great, unbiased way of identifying useful patient groups. It *will* undoubtedly lead to better treatment decisions. We call this **precision** medicine. We call it that because it is *different* than what we currently do. Imprecision medicine, which is built on a whole bunch of compromises and simplifications that work *really well* despite it all.

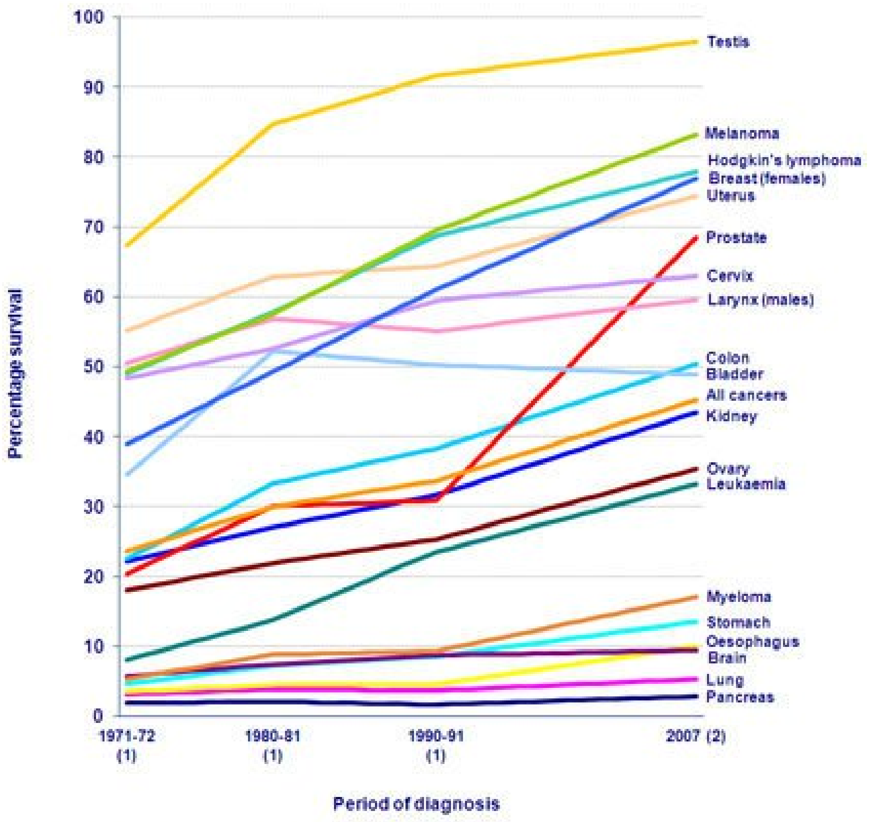


Figure 1.2: Ten year relative survival (%), adults (15-99 years), selected cancers, England and Wales: survival trends for selected cancers 1971-2007

(1) 1971-1991 Cohort analysis - actual survival
(2) 2007 Hybrid analysis - predicted survival

My favourite chart in medicine. We are winning, even without being able to do prediction.

The point is that we need a defined idea of what "beating doctors" actually looks like. If we accept that a machine outperforming a doctor at anything *vaguely* medical is enough, we have trivialised the entire concept. Self-driving cars are better than humans at driving without using hands. It verges on tautological.

Prediction isn't the only place this error in understanding rears its head. Look at the widely reported Autonomous Robot Surgeon Bests Humans in World First, where a robot "outperformed" human surgeons at suturing (stitching) up a pig's intestines. Again, amazing work by an amazing team in a very good journal. They created an autonomous bowel-suturing robot. This is a great step forward. In context.

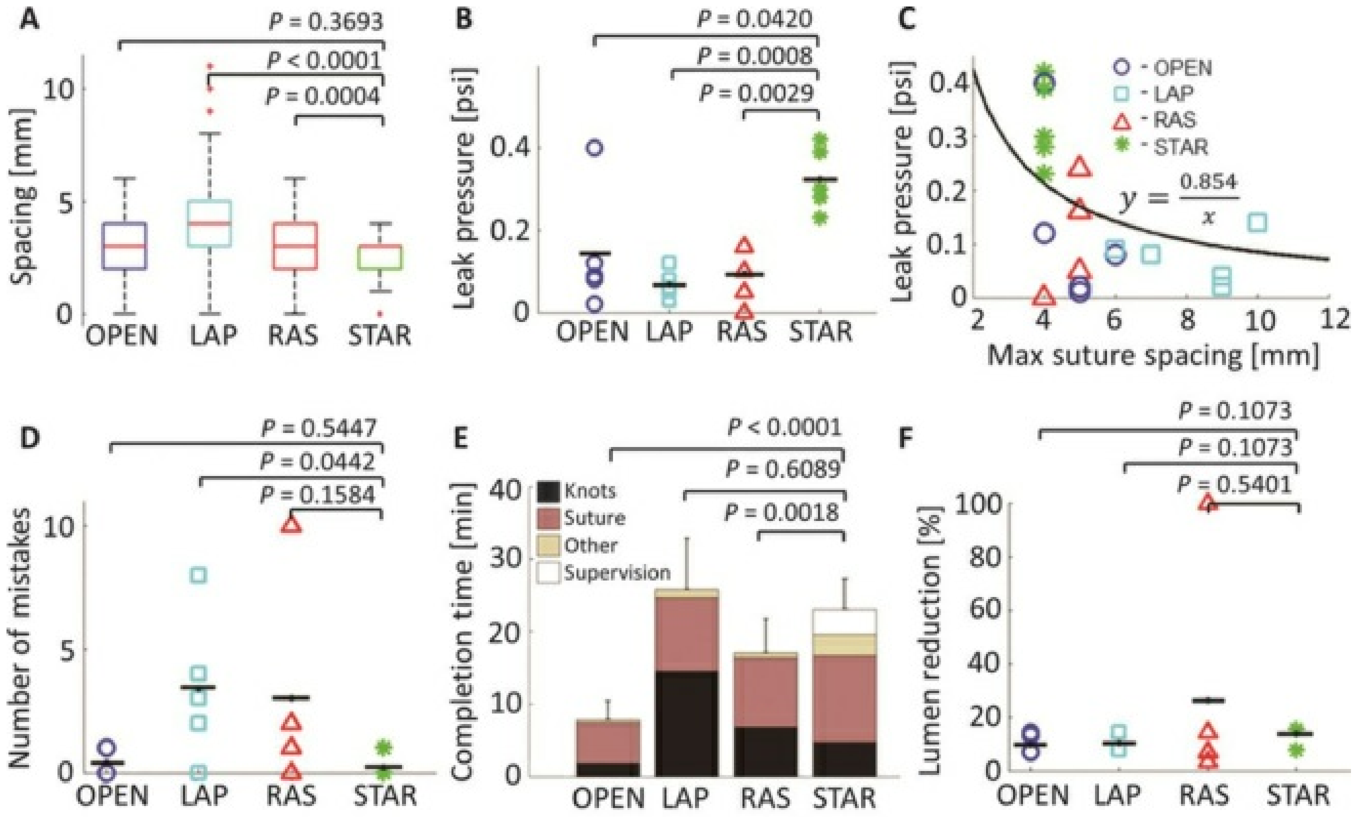


Figure 3 from the paper. Some really nice results.

Look at the figure. What did they test? Exactness of suture spacing. How much pressure it took to force the repaired bowel to leak. These are mechanical metrics, and it is unclear how they relate to outcomes. Leaking *sounds* clinical, but there is no proof that there is a direct relationship between the force needed for leaks, and the number of actual leaks in practice. There could be a threshold effect, and there is no appreciable benefit to "better" suturing. There could be a sigmoidal pattern, or other more complex relationship. Stronger anastamoses could be worse, stranger things happen in medicine (stents impregnated with anti-clotting agents created more clots). We just don't know.

The bottom three are different. Number of mistakes, time in surgery, presence of surgical complications. These are things that matter, that surgeons keep track of as metrics of their own performance. And STAR is no better than comparisons here. Much longer theatre time, no significant difference in mistakes or complications.

You are probably a little confused now, because it seems like I just described a different set of plots. STAR looks like it does pretty well in those last three.

STAR is *open* surgery. A surgeon would immediately understand this, and ignore the LAP and RAS results. It isn't a fair comparison. They cut a big hole in the pig, and pulled the intestines out through it to repair them. That is a big deal, and comparing it to a human using a laparoscope is like asking them to tie a hand behind their back. The risk to the patient is much higher with an open procedure.

We use laparascopic surgery despite slightly higher complications rates because it is better for the patient if *you don't cut a big, dangerous hole in them.*

Compared to human surgeons using an OPEN technique … STAR underperforms. Three times as long under general anaesthesia is no small thing.

As Andrej Karpathy says – human accuracy is not a point, but a curve. We trade off accuracy against effort. Surgeons don't bother with millimetre exact stitch spacing, presumably because it doesn't help. I'm not up to date on the last hundred years of surgical research, but I am totally happy to take as given that if more careful suturing helped, surgeons would be doing it (or maybe not, often culture trumps evidence).

It is the same thing with predicting cancer survival. Pathologists don't try to divide people into a dozen survival categories, if all clinical doctors want is to make a binary decision about surgery. It would be overkill. We do what we need to, and no more, and it already costs too much.

So maybe there is a more general rule for deciding when machines beat doctors?

Rule 1: Use a fair comparison

Rule 1a: Doctors don't do prediction

Rule 1b: Ask a doctor what they actually do, and what a fair test might be. Doctors trade off accuracy for effort, and optimise for outcomes (be it health, financial, political, cultural etc.)

Does this mean we need to do large randomised control trials to find out if any system actually helps with outcomes?

I wouldn't go that far. There are certainly tasks I can think of where the causal chain is understood enough to make an accurate inference. For example, in the paper above, luminal reduction post bowel repair has been tested thoroughly enough to know that a 20% or more reduction is needed to have a high chance of symptoms. We can use that as a comparison point. But saying 13% is better than 17% … we might need further testing to make that claim (or ask a bowel surgeon!).

So that is the first problem I see in "superhuman" medical systems research. But not all tasks are inappropriately chosen. Some tasks are exactly what doctors do, and we know exactly what doing better would look like. For example, Computer Program Beats Doctors at Distinguishing Brain Tumors from Radiation Changes shows that computers can do better than radiologists at distinguishing radiation necrosis (something that happens after radiotherapy) from brain tumour recurrence. This is very important, very hard for radiologists, and a great target for computational approaches.

Which brings us to the second common error.

---

2) These are not the AIs you are looking for

AI is AI, right? Machine learning is eating the world? Deep learning is so hot right now? Sure, except when it isn't.

Not all machine learning is created equal, and not all of it is groundbreaking, even if most people don't see the difference or think that it matters.

It matters.

Because the paper in AJNR (again, great paper, important paper) about brain tumours doesn't use deep learning. This is incredibly common in the radiology literature right now, because some major papers starting in 2010/2011 showed that an old style of image analysis could do some interesting things, like identify tumour subtypes in cancer cases from medical images.

These techniques are not loosely based on the human brain. They don't "see" the world. They aren't "cognitive" or "intelligent" or whatever other buzzwords are flying around.

These techniques have been around for decades, and we have had the computational power to run them *on laptops* for almost as long. There has been no hard barrier to doing this work for a long time. So why would it suddenly succeed now, when hundreds or thousands of previous attempts have failed?

Now, that isn't an argument on its own, but it should be *concerning*. Non-deep systems don't exactly have a track record of beating humans at human-like tasks.

The same techniques *didn't* beat humans in object recognition. They didn't help solve Go, or Atari. They didn't beat human transcriptionists or drive cars safely and autonomously for hundreds of millions of miles. They never left the parking lot.

Rule 2: Deep learning doesn't use human-designed features

The old style of image analysis was to get humans to try to describe images with maths, in hand-crafted matrices of numbers. This is super hard, so the best we could do is identify the building blocks of images. Things like edges and small patterns. We could then quantify how much of each pattern was in an image or image region.

This is what they do in the paper.

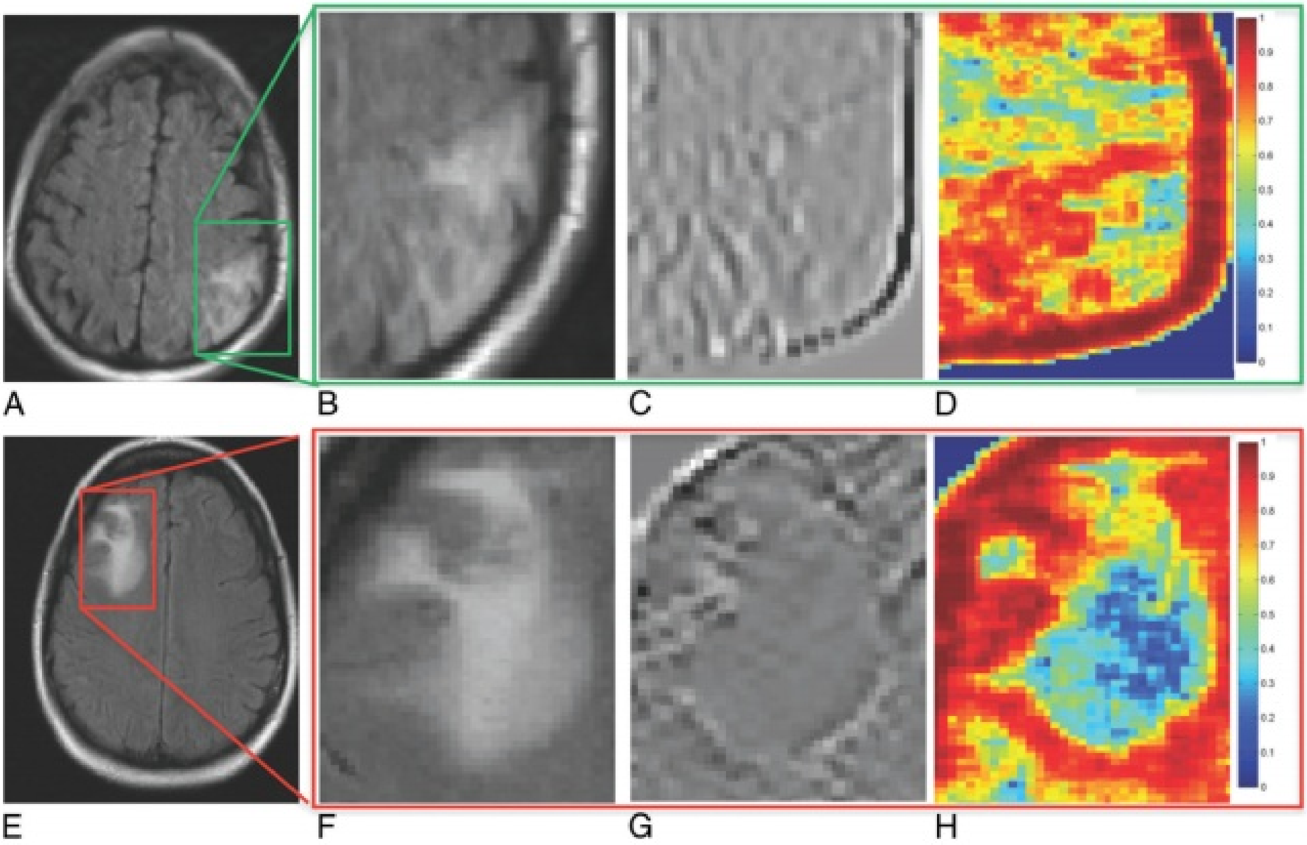


FIG 1. A representative 2D FLAIR section for RN (*A*) and tumor recurrence (*E*) shown for 2 different primary brain tumor studies. *B* and *F*, The original FLAIR images corresponding to RN (*A*) and tumor recurrence (*E*). *C*, *D*, *G*, and *H*, The top 2 texture features corresponding to RN (*A*) and tumor recurrence (*E*), respectively. Red represents high feature value, while blue represents a low feature value for a given pixel.

For starters, you can see why this is so hard for radiologists. A and E look identical.

What they are doing here is taking the region that is brighter (has more fluid in it) and quantifying how much of various textures is present. They try over a hundred textures in a cohort of around fifty patients, select the best performing ones and combine them into a predictive signature. They then use that signature to outperform humans *to some level of statistical certainty*.

Any statistically trained person reading this is hearing alarm bells right now, hopefully.

Probably the biggest problem with using human-defined features is that you will need to test them all, and select the best ones.

Multiple hypothesis testing is a weird beast. I really want to do a blog post on this at some point, because I really do find it strange. But the moral of the story is, if you test lots of hypotheses ("*texture x detects cancer*" is one such hypothesis), then you get false positives.

Feature selection – choosing the best performing features – probably makes it worse, not better. You expected twenty dodgy results, and you picked the top ten features.

I love the mRMR algorithm used in this paper for feature selection, and use it myself. But dimensionality reduction right at the end isn't a fix for overfitting. You've already overfit your data. The feature selection helps us explore the predictions and present them, nothing more and nothing less.

The thing is, all researchers understand this. We know that we are probably overfitting when we have very small n and larger p (small sample, more features than samples). We try to mitigate this as best we can, with techniques like hold-out validation sets, cross-validation and so on. This team did all of that. Absolutely perfectly, it is quality work.

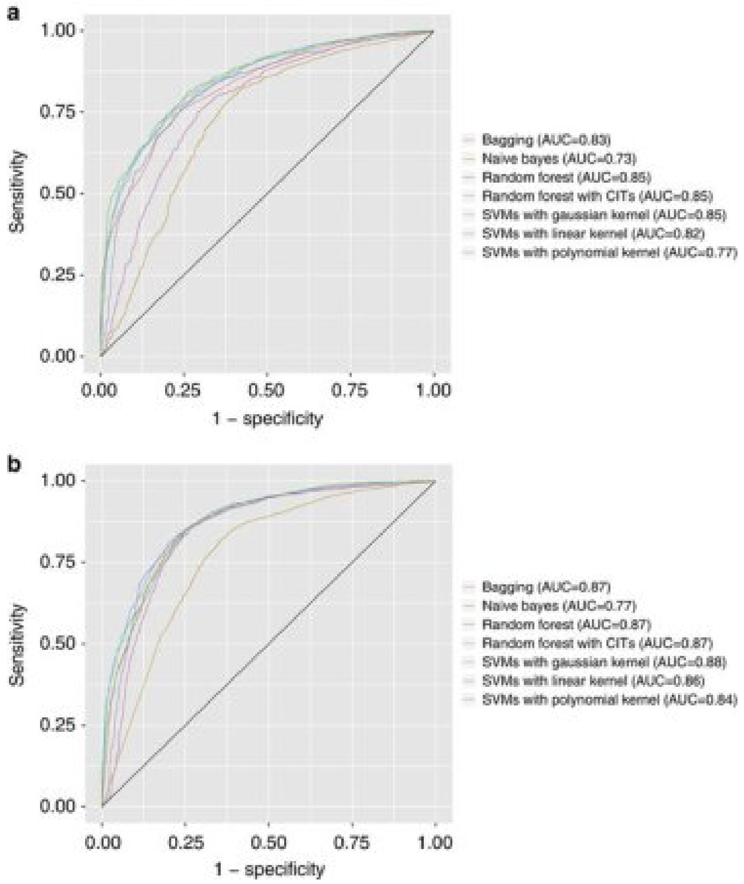
But all researchers in this field still know results like this can't be trusted. Not really. We might not need large randomised clinical trials, but unless a system is tested on a lot more cases, hopefully from a completely different patient cohort, forget about it.

But don't take my word for it. Let's read the paper.

> *Our study did have its limitations.* ***As a feasibility study****, the reported results are* ***preliminary*** *because our study was* ***limited*** *by a relatively small sample size, both for the training and holdout cohorts.*

Emphasis mine. The researchers are exactly spot-on here (I honestly can't think of an example where medical researchers of this calibre have overstated their results).

It isn't just sample sizes. You can perfectly split your train and test sets, but if you try a dozen different algorithms to see which one works best, you have overfit your data (picture from the Stanford paper again). Which is fine, again, but needs to be recognised.



Testing multiple algorithms can tell you a rough range of the true test accuracy, but you shouldn't expect the same results in a new data set.

One more thing to say here, a little more controversial. Public datasets. Be very cautious with public datasets, especially if you have worked with them before or have ever read a paper or blogpost or tweet about someone else working on them. Because you just contaminated your test set, Kagglers. You know what techniques work better than others in *this* dataset, which has its own idiosyncrasies and biases. The chance of spurious results that fit the bias rather than the true research target is very high.

Many machine learning researchers feel this way about ImageNet, and don't get very excited by the weekly "new state of the art" results unless there is a big jump in accuracy. Because hundreds of groups are working on that data, trying hundreds of models with wide hyperparameter searches. There is no chance they are not overfitting.

My machine learning colleagues shrug their shoulders. It is just accepted, take each result with a grain of salt and move on. It would just be nice if someone told the journalists and the public.

So, a better formulation for rule 2.

Rule 2: Read the paper

Rule 2a: If it isn't deep learning, it probably isn't better than a doctor.

Rule 2b: Overfitting is easy and unavoidable in small and public datasets. Look for larger scale tests, multiple unrelated cohorts, real-world patients.

---

3) That doesn't mean what you think it means

Whew. Almost there, thanks for sticking around this long.

Type 3 error is easy. The article never even mentions what the headline states, or the article completely misunderstands the research.

Digital Diagnosis: Intelligent Machines Do a Better Job Than Humans from Singularity Hub is a good example. There is not a single mention in the article of a head-to-head comparison. It is all projection and conjecture. It isn't necessarily a bad article, even. But the headline doesn't fit.

Artificial Intelligence Reads Mammograms With 99% Accuracy from Futurism is a bit more egregious. This article is about research in using natural language processing. It has nothing to do with reading mammograms, but instead mining the *text* of the reports that radiologists make. The headline is wrong, and so is a lot of the article.

Rule 3: read the article

Easy peasy.

The doctor is victorious…

So where does that leave us?

I remain convinced that we have yet to see a machine outperform a doctor in any task that is relevant to actual medical practice. The slowly building wealth of preliminary research suggests that won't last forever, but for now I haven't seen a case where the robots win.

I hope my rules will be useful, to help distinguish between great research that isn't quite there yet, and the true breakthroughs that are worth getting very excited about.

And if I have missed a piece of research somewhere, let me know.

…

…

Except, while I was writing this – literally this last paragraph – **it became untrue**.

Google just published this paper in the Journal of the American Medical Association (impact factor 37). And since it actually lives up to the hype, it is a great way to end this piece. Because any worthy set of rules should still work when the situation changes.

They trained a deep learning system to diagnose diabetic retinopathy (damage to the blood vessels in the eye) from a picture of the retina. This is a task that ophthalmologists currently perform using the exact same technique, looking at the retina through a fundoscope.

Google’s system performed on par with the experts, in a large clinical dataset (130,000 patients). While this isn’t necessarily “outperforming” human doctors, it probably costs under a cent per patient to run the model. An opthalmologist costs a lot more than that, and honestly has better things to do with their time. I am happy to call that a win for the machines.

Let’s look at my rules. Do they work?

Rule 1 – is it a task human doctors do, done with the same inputs. Yep.

Rule 2 – is it deep learning, with a decent sized dataset. Yep.

Rule 3 – is it actually a thing? Yep.

So you see, I can be proven wrong with my own system. Science!

Can’t call me a cynical, turf-protecting doctor now.

As a final note, it is worth looking at why the Google system worked. They paid to make a good dataset. *A lot*, presumably. They had a panel of between 2 and 7 ophthalmologists grade *every single one* of the 130,000+ images (from a set of 54 ophthalmologists). That is a huge undertaking. I don’t even *know* 54 ophthalmologists.

This technology is probably close to ready for a large randomised control trial, and that is a HUGE deal.

This is what we will see in the next few years. There will be many tasks like this, where computers can do exactly what humans do *if* someone is willing to build the dataset. Most medical tasks probably aren’t right for it, but enough will be that this will start to happen frequently.

Exciting times indeed.