2021 Podcast Series

Podcast 1. Hannah Guest. "Should we believe our ears? Questionable research practices in hearing science."

This text is an edited transcript of a recorded podcast.

Hello and welcome once again to a ManCAD / British Academy of Audiology podcast. You might well know that ManCAD stands for Manchester Centre for Audiology and Deafness and that we are located at the University of Manchester in the UK.

I am Gabrielle (Gaby) Saunders. I'm a Senior Research Fellow at ManCAD and I moderate these podcasts.

We always try to address the topics pertinent to the practice of audiology but also want to make sure that they are relevant to researchers and anyone interested in hearing and hearing loss. Some of them are specific COVID related issues and others are more general considerations in audiology.

We will record a new podcast each month each one will be about 20-30 minutes long and we will post the audio recording along with a transcript on our University of Manchester webpages.

You can find information on the ManCAD website.
http://research.bmh.manchester.ac.uk/ManCAD/Podcast/

Today's guest is Hannah Guest and will be discussing "Should we believe our ears? Questionable research practices in hearing science". Hannah is a Research Associate at ManCAD.

Hannah: Hi thanks for having me. I am, as you say, a Research Associate. I finished my PhD in 2018, so am still at the stage of figuring out my place in auditory research. In general, my work has tended to look at ways of measuring and understanding hearing differences beyond the audiogram. Ways that the cochlea and auditory brain might be affected by noise exposure or ageing without it showing up in audiometric thresholds, and how hearing might differ in neuro-divergent populations.

Gaby: Today we are going to be talking about how scientists and researchers in science should take care when interpreting research data, with some warnings about how to avoid falling into the trap of over-interpreting data, i.e. mistakenly believing that findings are saying more than they really are. Hannah, have I summarized that properly?

Hannah: Yes that's fairly spot on. The headline really is that science has gone a bit wrong somewhere along the line, and we as a field need to think about data more rigorously or else we risk believing in scientific phenomena that aren't real. This can have dire consequences in terms of wasted research money, people wasting large portions of their careers looking for effects that just are not real, potentially even consequences for patients and the public.

Gaby: What led you to start thinking about this in the first place?

Hannah: Two things had to happen. One was reading about solid evidence of widespread problems in science, especially fields similar to our own. The other was seeing those problems happening closer to home.

By the time I was midway through my PhD, there were already important papers on widespread problems in biomedical and psychological sciences. In particular, issue of important results not being replicable. That means a research team carry out study, produces evidence of some relation between things e.g. some supplement and bone health. That study is repeated with same methods but a different sample but produces a different result second time round. The crux of the issue is in many cases, perhaps most cases, the reason a finding can't be replicated is that wasn't real in the first place. Therefore, although this state of affairs often referred to as "replication crisis", really it's a falseness crisis. Because the purpose of science is not replicate findings, but to make claims that are true.

At the time I was reading about this stuff it felt like a faraway problem, because the people I worked with seemed to be really careful about avoiding this stuff. It's only been as time has gone on and I have seen examples in the real world and gotten better at spotting them that I realised we ALL need to consider. Not just junior researchers engaged in day-to-day; also anyone supervising/advising/managing researchers, and readers of research, and peer reviewers, and journal editors, and funders, and hiring committees. Anyone involved in research ecosystem has a part to play.

Gaby: You mentioned seeing these problems closer to home. What have you seen in the world of hearing science that makes you concerned for our field?

Hannah: Pretty much every corner of the biological and social sciences is affected by this stuff, so it's highly unlikely we're immune. Just from being an active member of our field, it's clear that we're not immune. Our research literature is packed full of studies with small sample sizes and low power, which should mean we see a lot of negative results, yet most papers report positive results. When you look at the p-values being reported, they show this unnatural pattern, with lots reported at 0.03 and 0.04 but almost none at 0.05 and 0.06. If researchers were not being tempted to nudge results in the right direction, this disparity would not exist. I can't remember the last time I saw a p-value of 0.05, at least not as the main result reported in a paper. Another thing I have seen that raises questions, I also see widespread use in hearing science of the Questionable Research Practices (QRPs) that we will discuss in a minute.

I would admit if you're a sceptic of the idea that there's any problem in our field; none of this is a smoking gun. Probably we need some kind of large-scale analysis of the literature to check how we're doing in hearing science. Perhaps, statistical analysis based on text-mining of a large number of papers. I think that's probably what it'll take to get a critical mass of people in our field to take this seriously, making sure we all do better work in future.

Gaby: Interesting approach. Hopefully by the end of your podcast we all go away and rethink what we do on a daily basis. What do we think is causing these false-positive results?

Hannah: False positive results, claiming to find a relation between two variables when that relation doesn't exist. The answer to that has a few different parts: **what drives** the process (incentives and structures pushing researchers to do this stuff) and **how** the false results are generated (mechanisms). The first question is probably the most interesting, and arguably the most important.

Throughout our careers, researchers are rewarded for positive findings. Academia is competitive, far more people with PhDs than there are permanent academic positions, so researchers need to stand out to survive. You need to publish a lot and appear ground-breaking. Most journals really like publishing positive results, especially novel positive results. From a cynical careerist point of view, you would conclude it is more important to generate novel positive findings than it is to do good, rigorous science. Currently, if you want to make it as a researcher, dodgy methods get results!

Gaby: "Dodgy methods", what are we talking about?

Hannah: More formally, they are what we commonly call questionable research practices (QRPs).

A common one is P-hacking which is altering your analysis methods, maybe many times, until you get a positive result.

HARKing (hypothesising after results are known) testing for relations between lots of different variables in a study, and then when one test comes up positive, act as if that's what you hypothesised in the first place.

Opportunistic stopping, waiting until you have just enough participants to show a positive result, possibly by chance, and stop recruiting in case adding extra participants would make the effect disappear.

Creative outlier removal where you decide that some data points are outliers because you have seen that removing them makes your analysis go in the right direction; or decide that potential outliers aren't outliers because leaving them in makes your analysis go in the right direction.

Failing to correct for multiple comparisons. When you run a lot of statistical tests and fail to adjust the required p-values, which is something you need to do when more than one statistical test. Or maybe you correct for some multiple comparisons, so you can say you've done it, but actually there are a bunch of other hidden comparisons you haven't corrected for.

There are probably other QRP's I've forgotten, but what most of them have in common is that they rely on making decisions about the design of your study when the data are already in – "post hoc" decision making. Many of these QRPs are made possible or made worse by small sample sizes. Those two points are important to bear in mind when thinking about what we can do to fix what's broken in hearing science.

Gaby: There's a lot there, I have a whole bunch of questions. Are you suggesting that scientists are deliberately producing false positive findings or is it happening 'accidentally'?

Hannah: I think it's probably an interesting combination of the two. A lot of it is unconscious or not fully conscious. The seductive thing about QRPs is they can get you positive results without you necessarily feeling like you're really doing anything wrong. I'll give you some examples:

With opportunistic stopping. You're running a study, and you're not absolutely concrete about how many participants will be recruited. As you go along you run interim analyses and each new participant changes the analysis results slightly. You get to stage where you see the effect you want, and you quickly stop collecting data. Maybe if you had carried on the effect may have disappeared or reversed – maybe this "positive" result is just a random statistical variation. However, what you can tell yourself, "this was the natural point to draw study to a close. I need to stop and get this written up before the summer, and 25 is a pretty good number, and probably is what I was aiming for!"

Something similar happens when it comes to P-hacking. Lot of different flavours, but means you tweak your analysis until it gives a positive result, and then report ***that*** version of the analysis and ignore the rest. P-hacking can be especially easy and effective in auditory neuroscience, because there so many different ways to process and analyse brain data, and you can tweak all of them at the same time until you find the magic combination that gets you the result you want. Again, as a researcher, probably not clear that this is what you're doing. You are just playing around with the data and when that positive result appears, you think, "Oh, well I just was doing the analysis wrong until now. If I'd really thought hard about this in advance, ***of course*** I'd have quantified my waves in this particular way. And ***of course*** I'd have chosen these particular artefact rejection criteria. And ***of course*** I'd have controlled for that confounding variable but not controlled for that other one."

Then, with HARKing similar story. You've got a data set with loads of variables that could plausibly be related, explore data by informally running some analyses. When variable C shows a relation to variable F, you think, "Yeah, I'm pretty sure that is what I predicted".

I may be horribly naïve, but I believe that the more researchers know about these "ways to cheat", the less likely they are to practice them. I don't think that most researchers are sociopaths using QRPs deliberately – they're sleepwalking into it. So the more conscious we are of all the different ways they could force a positive result, the more vigilant we can be to avoid these pitfalls.

Gaby: So people are not deliberately falsifying their data. They are playing with it until they find something and are just so excited that they run with it?

Hannah: I think that's absolutely right. I think most of us are wandering into this without really being aware what we are doing.

Gaby: I was thinking about not publishing negative findings or null findings. I remember thinking about designing an experiment, I happened to say to someone, this is your field, have you tried this, I was thinking of doing this. Their response was tried that, didn't work. I didn't do that experiment because I had that conversation, but I thought if they had published that, how much research do people do that someone has already done and said it didn't work but never saw the light of day so nobody knew it didn't happen or didn't work.

Hannah: This is such a huge problem. One of the many devastating consequences of us failing to publish negative results. One of them is the massive amount of wasted research time with people trying something that's already failed. There has been some smart modelling studies that have shown that publication of negative results is the single most important factor in making sure that false claims don't become canonised as facts so preventing a false claim becoming so widely accepted that we stop questioning it anymore, which obviously is a disaster for scientific knowledge. So it needs to happen. There absolutely needs to be publication of negative results. A failure to do so is not a benign thing.

Gaby: I guess that comes down to the Institutional issue. It can be hard to get a paper with null results accepted in a Journal.

Hannah: That's something I am going to touch on later and potential solutions. That's right up there.

Gaby: Let's lead onto that. What can we do to limit or eliminate these questionable research practices?

Hannah: Based on what we have just been talking about. Personal responsibility by researchers is important. Having the awareness of QRP's that allow us to do so. Large-scale institutional change is needed as you have alluded to. This is needed to alter the incentives and pressures on researchers. When researchers are being hired or promoted, we need academic employers to look at quality and rigor of a person's research, not just publication and citation counts. Journals to be welcoming of replications, welcoming of negative results, enthusiastically welcoming I would say and this apart from anything else

puts less pressure to churn out positive and novel findings. Need to create structures that help researchers be rigorous about avoiding QRPs. Researchers need to be forcefully encouraged to lock down study designs in absolute detail before gathering data, so there's just isn't any scope later to P-hack or HARK once the data have been gathered. This is achieved using pre-registration, where researchers upload very detailed study protocols to somewhere like the OSF (open science framework) website. This means these protocols can be checked later to see if they did what they said they would do. Journals can also support this by insisting on thorough pre-registration for all confirmatory studies. They can also encouraging novel publishing approaches like 2-stage review and registered reports, which are all about judging studies on methods, not on whether they produced exciting-looking positive results.

Gaby: I do have a question about that. Presumably, one can discover things that one hadn't hypothesised about beforehand. How can we allow that to happen if we are only working to the framework we propose upfront?

Hannah: You are right. Vital that un-hypothesised events are explored and made public so that other people can look at them in more detail. The way we handle this is a framework where we distinguish between confirmatory analysis and exploratory analyses. When the data comes in, if you have made clear hypotheses about things and laid out in detail exactly how you are going to analyse the data to test these hypotheses then those are your confirmatory analyses, so these are your research questions that you are investigating to find out if there is solid evidence for this effect or not. Any other poking around that you do in your data has to be regarded as exploratory and when it comes to that, you are not really gaining evidence for that effect when you are doing those exploratory analyses. What you are doing is seeing if there's some hint of an effect which is then about generating new hypotheses for further studies. The crucial thing is that when researchers are reporting the results of a study they need to distinguish explicitly between which are their confirmatory analyses, which should have been set out in a pre-registration protocol, and which are their exploratory analyses that are more about generating new hypotheses for further research.

Gaby: What about on a small scale? Are there things that research groups like ManCAD should be doing?

Hannah: I think there is plenty we should be doing, and a lot of it is already being done. We need to insist on high-quality pre-registration and thankfully I think that's something that ManCAD has got under their belt. We need to be able to offer researchers good, critical statistical advice and training. Make sure every researcher knows in detail about QRPs, how damaging they are and how easy they are to fall into. On a more human level, we also need to be encouraging young researchers to get comfortable with negative or "messy" results, so they are not tempted to "fix" them. In fact, I would contend that bosses need to be celebrating when their underlings produce negative or messy results as evidence that they are working honestly! Also need to encourage replication studies. Strongly encourage publication of negative results. Also that we are resourcing adequately powered studies, because many QRPs are made easier or made worse by small sample sizes and low power.

Gaby: Two approaches to this, the ground up level which is us as researchers need to be checking our practices, and the higher level of institutional change which needs to come from above where journals are more accepting of negative results, institutions change their expectation and eventually we can improve things.

Hannah: I absolutely agree. Each one of those aspects is necessary but not sufficient to solve the replication crisis.

Gaby: Unless they happen together. What about as consumers of research, of us reading journal articles. How do we figure out which is trustworthy? What should we be doing?

Hannah: A few things. Keep an eye out for small sample sizes especially if coupled big claims! Keep an eye out for p-values just below 0.05 or just below 0.01 are a bit of a tell-tale sign. Think you need to keep an eye out for lack of correction for multiple comparisons or inadequate corrections for multiple comparisons, where something seems to have been done in that regard, but perhaps not everything that needed to be done. Take novel claims with a pinch of salt if not replicated. By far the biggest culprit is hidden multiplicity: hints that researchers could have chopped up and analysed their data set in many different ways; then picked the analyses that gave positive results. Red flags for this are studies with a large number of potential independent variables and dependent variables if they offer scope for a number of potential comparisons that haven't then been reported by the researchers. Hypotheses that look a little odd – not the obvious way to use the variables or address the research questions, got to have an eye out there for potential HARKing.

But, a lot of this potential post hoc manipulation is hard to spot, only real way is good, thorough pre-registration. So if there's one thing listeners take away from this conversation, let it be that science will probably stay broken until researchers get comfortable with pre-registration and get comfortable with doing it well.

Gaby: Good message. I know that I am guilty of many of these things that you have been talking about. It will really make me go away and think; I need to be a lot more clear. I think that's what we need to be clear and direct and accepting of null results.

Hannah: Clean, direct and straightforward and actually just honest about how easy this stuff is to do and how probably very few of us who have absolutely clean hands. I am sure I haven't at all times. I think the more we get honest about our failings the easier it is for us to address them.

Gaby: I agree. Hannah, thank you. Very thought provoking and valuable. I hope many people listen to this and reconsider some of their practices. As you say, it's not directly trying to falsify results, it's just being swept away in the approach that people take these days.

Hannah: Thanks Gaby for having me. My email is Hannah.Guest@manchester.ac.uk if anyone wants to contact me about these or related issues.

All it remains for me to do is say thank you for your time and sharing your thoughts. Thank you for this.

If the audience have any follow up questions, feedback or share ideas for future topics please contact me. You can send me an email. Gabrielle.Saunders@manchester.ac.uk

I hope you enjoyed this discussion and are going to come back to the next podcast. Until then farewell and stay well.