NTDS_011

Key:

I: Interviewer
R: Respondent

I: Thanks again for agreeing to be interviewed. So, basically, I would like to start by asking you to introduce yourself and your background and the kinds of research projects you have been involved with, so then we get an idea of how to compare them.

R: I'm currently a [redacted] doctor [redacted] So I've used sale data to look at [redacted] questions, really. So I've been working with an analyst and the analyst has been getting the raw data and then I've been doing the statistical analysis.

I: So you're using also other kinds of databases and data infrastructures. What are these?

R: Simple databases for a lot of the clinical research, but then bigger datasets for the genetics, so whole genome and whole exome sequencing sets. I've been working with those datasets of writing my own pipelines to analyse and fit into them. So that's another area of the research.

I: There exist some databases for sharing genomic sequencing data?

R: There are a lot of databases for sharing certain types of data from sequencing sets, yes. So I've been using those more than uploading data to them or creating...

I: Which databases have you been using?

R: There are a lot of the NCBI ones which I've been using, Gene, dbSNP and then there's a new one, exaC, which is the exome consortium. So things like that are useful when looking at genetic results. There are a whole host of others but the NCBI ones, obviously the ensemble ones as well. Then recently exaC's been quite useful, actually.

I: How did the sale projects come about and how did you then start to collaborate?

R: As part of [redacted] we funded our analysts' time just to see what we could do, really, because this was at the time when sale was just becoming available.

I: This was back in?

R: This would have been 2010 or 2011, something like that. [redacted] So we were looking at just a new thing really, [redacted] So I came up with a few clinical questions and also there'd been some research by people before me, [redacted] to raise what are the uncertainties and what are the research questions that need to be answered. So there's a whole host of hundreds of questions on there that were identified. As a natural follow-on from there, it was used to answer some of those questions.

I: How did you work out to find what questions to answer?

R: It was just a bit random, really, but based on my knowledge of what's out [redacted] and what is possible. A bit of discussion [redacted] on what's possible within the datasets at the time. Obviously sale is changing and the data's changing, but what's possible. So it's mainly based around the GP dataset and what you could do with the GP dataset, [redacted] and what sort of novelty the sale dataset gives you as a research tool. Because obviously it's got its disadvantages, but how could you use it to...

I: Was it clear to you what you could do with these datasets?

R: Not at the beginning. There was a bit of a learning curve [redacted] One of the main things is [redacted] cases. So we had to work quite a bit on that because you're relying on how GPs record diagnoses to identify your cases because you're looking at the data as anonymised. So that involved quite a bit of work and thought of how it's recorded and stuff, and developing an algorithm to define in lots of cases.

I: What were they key steps there to develop the algorithms?

R: Coming up with an initial idea and then an iterative process, really. So, for example, coming up with an initial idea of defining cases, then seeing what kind of prevalence and incidence figures that gave us. And also looking at the changes over-time to see if that was realistic or figures that we would expect based on other etymological studies and then going back and refining the algorithm if we thought that there was any sort of discrepancies. So we worked on that, came up with something that we're fairly happy with and seems to fit in with other studies. That's the key step, [redacted]

I: What were the parameters that you were looking at in looking for figures that could match...?

R: You know roughly [redacted] prevalence and incidents should be from other studies, so you want your figures to match that. So that was the main

criteria. Obviously changes with time was a big thing because there have been some key moments of GPs in terms of recording data, for example when the QOF Quality and Outcomes Framework was introduced in 2003, there was a financial incentive for GPs to record it. We could actually see that. At that time there was a peak or a change in the graph in terms of prevalence and incidents. Obviously that's not a real thing that just reflects how the data's recorded. So I think you've got to keep those kinds of things in your mind when you're working with these datasets. You've got to have knowledge of how the data is recorded, really. I think having worked in a GP surgery and knowing what it's like, that was a little bit of a help.

So I know that in a busy surgery you're not going to take the time to record a very detailed diagnosis [redacted] You might not say that this is [redacted] You might just record it [redacted] because of time constraints. As a GP that's all you're interested in. You can understand then the limitations of the data you have to work with, I suppose.

I: How were you aware about the changing practices in the recording?

R: Just as my knowledge of working as a doctor, really, so clinical experience and looking at the trends.

I: Was it very iterative, this process?

R: Yes, it is quite iterative. The flexibility in the sale system was really good, enabling us to go back and change algorithms a bit and come up with different...

I: Was it a process just between you and Aaron?

R: A bit of a larger team, yes. So Mark Reese and Mike Kerr were the senior academics we've been working with, so their guidance as well.

I: Was it different from compositional teams in other research projects? Was it different kinds of expertise?

R: In my limited experience of other types of research it's the same thing. You've got senior overview and the senior people haven't got the time to always spend on this real nitty-gritty, the minutiae of it, as an overview, and then you've got other people with specific skills. [redacted] the SQL skills for example, looking at other people in the team who've got etymological research, [redacted]. Like any other kind of research, really, you're looking for people with fresh ideas sometimes. Sometimes that can be very useful. You can look at the same thing all the time.

I: You don't have access to the sale data yourself?

R: No. So I've got access to the gateway. So we've changed how we work a bit. At the beginning I didn't have any access [redacted] export a file and then I'd work on that, but now I have access for the last few years to the sale gateway, which is easier.

I: Is it a browser or an interface?

R: Yes, it's a SQL device which allows you to look at the raw level data, really. So it makes it easier. Then you can do analysis within the data.

I: So you have also access yourself [redacted]

R: Not right at the beginning but in the last few years, yes. It makes it a lot easier.

I: Was that a barrier?

R: I think the whole infrastructure was developing as we were going on. So sale was quite new and they were developing how to grant access in terms of gateway accounts and things like that.

I: Sale obviously, for security confidentialities, is anonymised and secures the non... so then does this bring particular kinds of things to do or issues in the research process?

R: Well, it's similar to other clinical research. We have to have IGR pre-approval, the sale information government research panel approval, and that's similar to other research like ethical approval. So that can be a bit slow sometimes. Then obviously you have to work within the gateway, which slows your research down a little bit because it's not a perfect system. It's slow. It gives you a limited screen size. It cuts you off every now and again. Then you have to export the data to use and get approved by another sale analyst, but that's generally fairly quick. So that's generally done within 24 hours. So there are limitations but they do seem to work quite well within sale. We've got used to using them.

I: How does working through the gateway change the way you would work?

R: Well, it slows you down a little bit. I think it's very good in that you can work anywhere with a UB key. So that's good, so you're not limited to working here, but everything's just a bit slower, really. If you want to share the data or show the data to other people, for example, you have to make sure it's not raw level identifiable and export it, whereas otherwise you just email your collaborator and say, "Have a look at this, what do you think?" With this, if they haven't got sale gateway access, you have to make sure it's in the right format, apply for permission to export it. So it slows you down but not massively.

I: To access the gateway brings you basically behind some protection?

R: Yes. It's another layer. So there's individual project-level encryption now as well, so you're only allowed to look at data that's due to your project. So that's another new thing. So there are obviously different levels of encryption.

I: I'm learning a little bit also how the system works because of schedule I'm starting work with the data users first and I'm going to meet most of the sale team in a couple of weeks' time. That's why I'm asking these kinds of questions because I need to get my head around it. I was reading the paper [redacted]

for issues of potential de-anonymisation. I was wondering if these were something that you had to do because of the sale licensing. Is it something that you would do in any other study?

R: Well, I think both reasons are true. There'd only be one or two people taking those drugs in that study, for example. So if you published a graph showing that just one person is taking that, then theoretically some people know who that person was. The GP would know, "Well I know I've only prescribed this drug. It must be this person." The person themselves, the family might not, the doctors involved. So there's a theoretical risk of identifying people, but also with that study, actually statistically it didn't make sense to include those because you haven't got... one or two people... it didn't give you the power. But reviewers do latch on to that and they say, "Well, this impairs the quality of the data," which it can do sometimes as you obviously put a bit of bias on it.

I: What are the strengths that you find most interesting about the sale data? One that I found that I liked was the same paper [redacted] said there is freedom from pharmaceutical industry bias because it's routine data.

R: It's routine data. That's the thing. It's interesting because in similar studies [redacted] the trial data doesn't come up with any side effects, but actually when you start using them in real-life then side effects appear and you think, "Why is that?" It's routinely collected. So it's free from that one particular type of bias. It's counteracted by the fact that the weight, for example, ideally you'd like to have everyone's weight checked the same time, like a week starting before and then two weeks after, or what have you. So a bit of noise is introduced to the data, but one of the major strengths, apart from the numbers, is that it is routinely-collected and it does reduce some bias definitely.

I: What kind of considerations did you have... all this data is the routine data, so it's collected by lots of different people with lots of different training? So I guess the algorithm was one way to clean or refine.

R: The volume cleans itself. There's a lot of noise being introduced with different people collecting weights, for example, in different ways. Like with anything, if you get the big enough numbers which sale can offer, you'd never be able to do a clinical trial. You would be, but it would be very difficult to do a clinical trial with the same numbers. It would have to be a multi centre UK trial and cost millions of pounds, basically. So that's a strength, but that's counteracted by the noise that's introduced with you not being sure, for example, how those weights were taken. You haven't set a protocol saying, "You collect the weight using this type of weighing scales and this time with this type of person," etc. So there are weaknesses. It's sort of random because you're collecting everything. You're not selecting people for a trial as well, so that's another strength. People can refuse to be part of the trial. There's ascertainment by trials as well, which you don't get in this type of analysis, because you'll include everyone. They can't refuse to be in.

I: Are you planning to keep working with sale?

R: Yes, hopefully. I think things are changing in terms of sale and funding is obviously an issue. There are more people doing similar types of analysis as

well. We need to make our research a bit more unique now and focus on the strengths of sale, because sale loses out in terms of numbers, things like the English GP database, because the England population's a lot bigger. You have to play to sale's strengths and its flexibility, ability to upload a lot of different datasets, the way the gateway is set-up, you can go back and do things which you can't do in other similar systems. You can only apply for one sort of data out and that's it, you get it.

I: Can you also initiate, if you've got access to a database, for example, and bring it to sale?

R: Yes. So that's a strength as well. That's one of the things we're doing at the moment. We're uploading our own data to sale, getting it anonymised and linked using the...

I: Is it one of the databases that is on the table from the website?

R: No, it's different. [redacted] We're doing a bit more rigorous thing now to validate our case ascertainment algorithm. So we've collected people we definitely know [redacted] uploaded to the sale and then we'll link back to the GP and we'll be able to work out the precision and recall of our algorithm a bit more definitely, which has always been a criticism of our work. How do you know you're doing it exactly?

I: Then that data stays in sale?

R: Yes. There are some rules and ways that they change access, but yes, this stays in sale.

I: So you make it available to them?

R: Yes.

I: So it's not just comparing for the study and then...?

R: No, and if other people wanted to use it, it's anonymised so that shouldn't be a problem for other people wanting to use it as well. That's the next big step for us, to upload more [redacted]

I: Are there other strengths of sale that is important for... the competition is actually growing.

R: The infrastructure's very good. The variety of data is very good. The people, (unclear 0.23.43), are leaders in their field and that's attracted a lot of good people. So it's good to have the people around you can work with, really. How are we doing for time?

I: I think you'll need to be going in a couple of minutes. Great. I think we were covering everything. Last thing, so now that you are acquiring experience of working with sale, also bringing the data that you wish to work with more into sale and comparing and stuff, but also maybe in the way you organise research or bid for grants and stuff, and how you assemble teams.

R: That's right. [redacted] So that is the next step, applying for grants and getting teams together, because that's where we're limited at the moment. We need funding for researchers, really, and now sale are introducing costs as well, so we need funding for that. So that's a new thing.

I: So it's a way to costing sale?

R: Yes. So now they are applying a cost to each project, which never used to exist. So that has to go into your grants, etc.

I: Okay. Great. Thanks a lot.

(End of recording)