NTDS_015

Key:

I: **Interviewer**
R: Respondent

I: **Yeah. Good.**

R: (Recording inaudible 0:00:04.1).

I: **Mmm?**

R: Sony versus Olympus.

I: **Sorry?**

R: Sony versus Olympus.

I: **Yeah, exactly. Comparison, and we'll see, if one doesn't work, then…**

R: The other one might.

I: **Exactly. The decision will be easy. Okay, so maybe to start with, we can just probably introduce your background, and the trajectory that led to work with SAIL?**

R: My full background was actually quite, quite different to the others. My first degree's engineering, and I did a master's in environmental technology, which brought me into geography, and I ended up doing a PhD on changes of climate in America, using satellite images.

I: **Oh wow.**

R: So basically I started working with large data sets during my PhD, so it was created data for North America, for (ph. 0.01.03) *half degree*, and looking at climate patterns in all of these little boxes.

I: **Amazing.**

R: And a job came up for SAIL. Yeah, I think… was it called SAIL then? It had probably just started to be SAIL, then.

I: **Aha.**

R: And which asked for someone who had experience with large data.

I: **Yeah.**

R: And I applied for it.

I: **That was immediately after your PhD?**

R: Yeah. I had just submitted it, basically. I applied for it the same week I was submitting my PhD. And that's how it started with working with this. And I think it asked for health background, and I had done a course on… I can't even know what it's… it's something like health methodology or something, an optional course during my engineering degree, which was all about dialysis systems, and what it meant for the patients, and a little bit of the biology behind it.

I: Right, yeah.

R: And a bit of clinical setting. And I thought, 'That sounds interesting. Let's give it a go.' [Laughs]

I: Yeah, absolutely. And so, then, it worked. You got into SAIL pretty early. So, what was then…

R: I think it was already running for a couple of years, but it was still very small then. It was just two rooms full of people, basically, at the time.

I: Aha. Amazing. Full of people? How many people?

R: Well, I think in my room it was a senior analyst, and another four analysts. There was a third room, but I never quite knew what those people were doing. I do not think they were doing work on SAIL, they were doing more the traditional questionnaire type of data. But they were linked to us. And then we had another room full of about six, six to eight people, who were the technical team at the time, and they were supporting and building up data structure, so at that time the SAIL gateway didn't exist – we had different methods, and we had to move over then to the SAIL gateway.

I: Right, okay, okay. And so you came in as a… what kind of role?

R: Research assistant.

I: Right, and-

R: So I started off.

I: Okay, what did you start off doing?

R: Lots of data… I think the first thing I had to do, for the first year, everyone had to write documentaries about… not documentaries, big documents about data sets, trying to describe all the variables, because we received the data, the documentation, we never got any official documentation with the data, so we had to write the documentation of 'This is where the data set comes from. These are the variables, and these codes mean this', basically.

I: And these documentation were intended for internal use, or…?

R: Mainly for us, but also if someone came onto the research project, we can refer to the documents.

I: And so why where you all doing this? You were acquiring a lot of data?

R: I think there was three or four. I think it was when we had gotten more new datasets in at the time, as well, and I think we maybe just realised that without at least some minimal documentation, we can't really do anything. If we don't know what 'Code 123' means, it's something called 'Birth Code 1234', how are we supposed to do research with it if we don't actually know what the variables are?

I: So these documents are now used as a…

R: Yeah, they're still used, and I think they're updated at the moment by Dan's team.

I: Right. And it's a way to get an understanding, a first approach to it, at the set, and to see what it contains?

R: Also, our co-researchers, we have a big team in Cardiff. They use it a lot to actually go through it and look at the variables they're interested in, so we're going to come back, and thinking, 'Oh, for our research, it might be interesting to know what this C-section code means. Can you have a look at how well it is populated, and even if we can use it?'

I: Right. Are these clinical researchers in Cardiff?

R: Yeah, clinical researchers.

I: And so, was it very challenging and time-consuming to write this documentation? Was it a lot of work?

R: It was quite a lot of work, but at the beginning we didn't have lots of other projects. I think now I would say we've basically had lots of time to do it. It was just a bit of a boring job to get on with. Now I think I would just be laughing, because now everyone has at least six projects at the same time to deal with. So, it is very different. I think they just… long time as well. If the computer went down, we couldn't do anything [laughs]. We just went and had a cup of a coffee. Now everyone would just get out their laptop, continue working.

I: Right. And so, what was the kind of problems that you were trying to solve by writing these documents? What could be the things you wanted to check, and…?

R: Well, I think we had to write the documentation because we were also starting to work at the time on the ways electronic (unclear 0.06.22) to children, which brings together lots of datasets. So basically, without knowing what the variables mean, we couldn't go forward to build this big platform for research.

I: So how did you know, then, what the variable means? You just had a look at the tables?

R: Well, one way was to contact what was then called NWIS – it's the NHS Wales Informatic Services. They are our third-party data providers, so the data goes through them. They strip all identifiable data from it, and they give us these… well, you've probably heard of the ALFs and the RALFs, the anonymised linking fields. And they have worked with the data much longer than us, because they do they official reports for the Welsh government.

I: Okay.

R: So they know where the data comes from, they know the data provider, so we ask them in the first instance to help us with all the codes. We say, 'What does this mean, and how do you use it?', and they gave us recommendations of 'This variable is not very well populated', and other things we found out ourselves. Quite interesting, the researchers in Cardiff wanted to use breastfeeding codes, and we found out that in south Wales they have recorded at a different time than in north Wales, so when they merged the datasets together, they found that we had a breastfeeding at birth variable flag, and the breastfeeding at four to six weeks variable flag, but south Wales, I think, only used one of them, and north Wales the other. So what they had to do for the research was join them together, which was not ideal, because the researchers would have wanted a better variable than it was, but it was not really quite what they wanted. So I think the main problem on the research side of things is that those are administrative datasets – they were never designed to be used for research, so they might not… they are potentially really good to do research with, but clinicians, if they're not particularly told to fill certain bits and pieces in on forms, it will never end up in the dataset. So you might have a variable that sounds really good, like breastfeeding, but no-one bothers to just tick it.

I: Right, yeah.

R: But to be fair, the data has changed dramatically. The data is getting better and better. So, you go back to the nineteen nineties – you have lots of missing data for variables, but now they're much better filled in.

I: Okay. And is that across all spectrum?

R: I think it's most datasets. I think we have the occasional glitches were someone finds something that doesn't make sense, and it goes a bit around in circles until someone figures out what happens. I think Dan would be able to tell you more about this – we recently had a big glitch in our GP dataset, and I think it just… one of the connections didn't upload data or something, so we're missing data from certain GP practices that didn't come through.

I: And so, how do you find out when you've got missing data, or missing variables, or…?

R: It depends. It's all a bit of trial and error. Getting to know the data sets, so basically we now have people who are really good and experienced in using certain datasets, and you just ask them to do it. In the beginning, it was just… either we found out ourselves by looking at the frequencies of what was filled in, or it got to the researchers, who said, 'Oh, wait a second, there's something wrong with this variable', or 'There's not enough in there. Could you check again?' So it's a very working-together approach, I have to say. You have to talk to other people.

I: Right, okay, yeah. So, you then, as you were talking to the researchers, so you were participating, am I right to say it was the role of the data analysts to write it, so that you were liaising between data analysts [overspeaking]

R: Yeah, yeah.

I: So…

R: So we were sitting in between. We have the technical team on one hand as well, so sometimes we had to refer back to the technical team, saying, 'There's something wrong. Can you please check?' The technical team would either then say 'It's nothing that's wrong in SAIL, this has come from NHS Wales Informatic Service. Something has gone wrong at their end, and they might have to re-encrypt the data.' So it's quite a long chain, and sometimes researchers on the other end might get a bit disheartened, because things might take longer along this chain.

I: Okay. And so, you talked to the technical person, and then you go back, you read through/engineer what could be the problem. So, then the Health Solutions Wales was your intermediary, to talk about the data sets?

R: Well, to some extent, they know more about the datasets than us, because certain people have worked with them. So, if you have questions about certain variables, or something like that, they're a very good contact point. The other thing they're contacted for is if we think something has gone wrong in the ALFing or RALFing process. And those other things would normally take a bit longer to sort out, because… and we would have to check the algorithm, and I think in a couple of instances they had to go back into the data provider, and we had a problem two years ago – they found out that from one refresh of the dataset to another refresh of the dataset, there were suddenly twenty thousand children less in the dataset. And it went through this chain until it came back to say that the local health board had archived all children at the age of eighteen, because after their definition, they were not children any more, so they removed them from their dataset.

I: Okay.

R: So in the end, I think, NWIS had to go to our last refresh, and merge them together with the new one, so we didn't lose children over the age of eighteen, because we still want to follow them up.

I: Okay. So, in this case, how did you find out? You just went to the data provider, basically? They explained [overspeaking]

R: Yeah, yeah. I think we approached NWIS first, and asked them if there's something wrong, if they can have overlooked something. It seems to not be right, and explained the problem to them, and I think they realised it was definitely not to do with what they've done to send to SAIL, but it had to do with differences of the data from the data provider.

I: The source.

R: And they, yeah, went back to the source.

I: Okay, okay. And instead, if there's a problem in the communication between you and Health Solution Wales, for example, you said the ALF and RALF, how do you find out? Because it's pretty complex, I guess, to get to the information.

R: Yeah, that can be… I've just been working on a project where, basically, the woman I'm working with, she knows very well the dataset that she prepared for us, and she looked at the year of birth, and the gender, and other values, and then she looked at now, on her project encrypted dataset, and she says, 'Well, they look different. Those can't be the same children. Something must have gone wrong when they were linked together', and then it has to go back and, and it needs to be [overspeaking]

I: Okay, so that was then… that's the same sort of story. Situation. It was then generating a cascade into the RALFing, ALFing and RALFing, so…

R: I think we had a problem with a RALF a while back… oh, that's probably my lunch call. I can ignore that. With a RALFs a while back, and I think we only found out because we looked at the RALFs on the map, and said, 'Wait a second, this is supposed to be just the Swansea area. Why are we getting RALFs in Carmarthenshire?' So we know something had gone wrong, because the area was wrong, and then it had to go back to Cardiff as well, for re-RALFing.

I: Right, yeah. Okay. Very interesting. So then, tell me, for example, more about the construction of the cohorts, especially the cohorts. You were saying there was lots of things to harmonise, and so far we've said you started by looking at the datasets, and then documenting, to get [overspeaking]

R: Yeah, and basically we had an outline from the researchers in Cardiff of which variables they wanted in the dataset, so you start off small – like you want the ALF linking against other datasets. You want a week of birth, a gender code, a date of death, a birth weight, a gestation age, and these things. But some of these variables exist in different datasets across SAIL. Basically, we found that the National Community Child Health dataset is the most trustworthy, because it's the hospital records, and so we deemed them to be more accurate for some of the base data. But then, the Office of National Statistics have probably been very correct as well, so we came up in the end – or one of my team came up with – some kind of precedence rule, saying 'If this is filled in, and this, the other field, is filled in, then take this field before the other field', and then the first time we just did it all by hand. The next time, we have re-run, we tried to automate it and start writing procedures that do it all for us. Basically we start off easier, ending up with an enormous amount of STML code, to try to code everything into the slots and pieces, and of course you get errors as well. And the next version, we tried to automate as much of what we've done before for the next step.

I: Right. So why was the first run… you did it all by hand? Every single children record?

R: No, no, I ran the children records, but we did all the harmonisation between datasets by hand, so we had a table one with all of them, and a table two comparing this and this, and a table three, comparing this field and another field. So it was just…

I: Running every SQL…

R: Yeah, running it all in stats, and so the next time round we tried to be a little bit more efficient in doing it, so we included a whole step of harmonisation in it.

I: And then you said, as you go on, you automate and you accumulate this experience. And where do you reuse it, this experience? As you are inputting new datasets, or is it also for other kinds of projects?

R: Well, basically a lot that we've done for (unclear 0.17.49), we then present it to others, and they said, 'Oh, this is really helpful, this methodology.' We've started all kinds of things. We said, 'There's not enough documentation done of how people do things. They basically just write SQL code and give the result to a research team, but what we need is a bit more documentation of what has been done. So we introduced the SAIL Wiki, as a forum where people can document the codes they use, so if I use a certain set of respiratory codes, I can put it on the wiki, so if someone wants to do another project, on children's respiratory health, they can re-use the codes. We have concept dictionaries saying, 'Oh, this is the concept – we want to identify this and this, and these are the codes we used', so trying to minimise the duplication of records. Of course, we found out that three or four people would do similar studies and everyone started from scratch, getting the clinical codes together.

I: Right [laughs]. And how did you find out?

R: Well, mostly just by chance.

I: How did it happen?

R: 'Oh, you just did a project on asthma. I didn't know that. I could have got the codes from you.'

I: Right. But was that because people were not really working… it's not that the people that were working in same room, or…?

R: Yeah. But also, I think at the time, it was just… especially when we moved. Was a year, two years ago when we actually moved into (unclear 0.19.09), who… we were already starting to split over different floors, and then the communication was getting much, much more difficult, with the other people, because…

I: Right. So, the solution has been the wiki?

R: Yeah. The wiki has been good. Not everyone is using it, but the ones that are using it are finding it really, really…

I: Helpful.

R: Helpful. Put something up if you find something. Also, if you have a problem with some software not working, instead of just reporting through the technical team, put it on the wiki and let the others know that something is not working.

I: Right, right. So then you upload a data script there, so it's also the library?

R: Yeah, so you can use it as a look-up library and template for other things.

I: Okay, great. And do you have also meetings between the analysts, where you can discuss, and be up to date?

R: Yeah. We have this SAIL user forum, and Dan's organising that, so we've been talking more about it. We're meeting once a month – I think that's for the internal analysts, and every quarter of the year we have one where all the external people are invited as well, and we might have bigger talks in more interesting subjects, and the little ones might be more about coding and other things that aren't more interesting to the other people who work with the database day to day.

I: Okay. So, when you were… if we move now to the way you work with the clinical researchers that get to know SAIL as well, so others that have interactions – how do you understand their requirements? How do you work out what is feasible to do? Things like that.

R: Normally, we try to get some kind of project description document off them. Sometimes it's difficult – some people are busy, and you end up with one hundred emails giving you bits of code, and thinking 'Oh, this should be added to it', so ideally you have one document that specifies the requirement for each project, and nowadays… I think in the beginning, it might not have perhaps been quite that way, but we always have clinicians involved as well, so if researchers say, 'I would quite look at these codes, and these other ones I've found', another clinician will say 'Is it sensible to use these codes?' We'll actually check over it as well. And then what we normally do is: We put all the codes into SQL, and look at a frequency distribution of how often they're actually used. Of course, what is special with GP data codes, there are codes there but GPs don't use them. A GP will only use codes if there's an incentive to do so, like the QOF criteria. The Quality and Outcomes Framework, something like that. So there's an incentive to code, for example, diabetes, because they get money for using the code. So every time they have someone with diabetes, they take their blood pressure, and they tick another box, they get money for their practice. So you end up with lots of codes not being used the way you think they might be used.

I: Right. So, because you've got an understanding that there is a certain distribution in the population, but it's not reflected in the data?

R: Yeah, it might be… you think there's this code, should be used for it, but the GP doesn't have time to search for the code, and he goes for a higher-level code. So instead of just having a proper code, for example, for diabetes type one or type two, they might only use diabetes, and you can't distinguish between the two groups, because only the top group is used. So it makes it a bit more complicated. We need to think a bit more about ways of how to detect people, so, for example, with asthma, we actually found more people through treatment steps. We had a code that looks at asthma reminder letters, or something like that.

I: Okay. Yeah, and that's [overspeaking]

R: Which is just an administrative [overspeaking]

I: And then you understand that they have got a diagnosis for asthma.

R: Yeah. So they have a diagnosis for asthma, and of course they send these letters.

I: Right. Where the record doesn't have a diagnosis code for the asthma diagnosis?

R: Yeah.

I: Okay, okay. Yeah, yeah.

R: So you end up getting very creative with the data. 'How can I detect these people, and why do I not get a certain population of epilepsy? Why don't I find it?' You need to figure out to… I don't know if you're talking to Martin, but Martin's very funny about… he says it's like detective work, because you really need to dig into it, and you need to… you say, 'Ooh, this is really interesting. What's happening with this? Why can't I find this person here?' [Overspeaking]

I: Yes, it's very…

R: But you have published figures that will tell you there's a certain incident rate of certain diseases, and you look at them, and go 'I can't find them. Where do I find them?' And then you compare different codes together, and the you work it out in detail.

I: Sure, yeah. Very interesting. So, then, what do you do? You find these asthma reminders, then I guess you prepare an algorithm to…?

R: You check it with the research team first. You say, 'I've found this. Do you think we should include this?' And then you just… it can come up. Normally they get quite project-specific, so you might have to have a very project-specific code for how you define asthma, depending on the research question they're asking.

I: So if that is sensible, it leads you to make the interpretation. And so then, you prepare algorithms to use to do that, so [overspeaking]

R: Yeah. You prepare the data for the research team.

I: Right, yeah. So then, do the research team use… you provide them this view, right? Which is the end product of this. So, for example, in the case of the asthmas, then you have recent… with them, through this way of extracting the asthma figures, and then you provide them with the view.

R: They normally want a flat data table for analysis. Just and SPSS file, or starter file, and they tell you, 'Oh, we want the ALF, and we want the week of birth, and the gender, and we might want the date of the first admission to hospital', and we give them all the different flags. Normally we have to do quite a bit of recoding as well. They say, 'Oh, the codes are as follows in the original data sets, but they're actually not quite correct, so we have to do all these case statements to recode the data to whatever the project team [overspeaking]

I: Right. So then these tables, these refined views, they've got also some derived data, sort of metadata that you have worked out.

R: Yeah, yeah. They will have derived, yeah, they will have derived that all, yeah.

I: Yeah, and then, do you keep that derived…? How does that get inherited? Does that get inherited through the library in the wiki, or do you also keep that data somewhere?

R: Normally, at this stage of the project, what we are having, is we are having… each project has a project number, and on our IBM DB2 databank, those will have been different schemas, so they start with something like 'SAIL0009V will be the schema for the original record project', and actually we have two now… in the beginning we didn't, but now we actually have two schemas for the project, so we have one that is the exploratory schema, so X-something, which is where we prepare all the data, and we have one without the X, which will be the tables already project-encrypted, for the researchers to look at. So we keep a copy of what we did, so we can compare it with other data sets if we get questions back, and then we have one copy which is a project-encrypted version for the researchers. And they stay there, so even if the project expires, we will lose the access to the projects, but the data's still there. So if we needed to go back in ten years' time, saying 'Oh, I did something like that, but I've lost all my code. Can I just have a look at the tables?' You can apply to get permission from Cynthia again, to get permissions for a couple of days to go back into your folder, and [overspeaking]

I: And that is the project-encrypted data, you mean? That one?

R: Yeah. Or the other one.

I: Or the other one.

R: Yeah. For the project, of the non-project encrypted data, depending on what you're looking at.

I: Okay, okay. So both these exploratory folders and project encrypted data are contained with the [overspeaking]

R: Yeah, they're containered and, yeah… and they're all backed up. I think there's a backup happening on… Rohan would probably know. On certain days, at certain times, they're backing up all the information on the databank, so we can't lose any of this data.

I: Okay, yeah.

R: And what we're now trying to do, I don't think everyone is doing it, but Dan and a couple of the analysts who've been along for a long time, we're using SVN version control for scripts as well, so we have a project folder which will say '009 (unclear 0.29.00) project', and then subfolders for SQL code, documentation, other things, and they will try to keep that all updated on an SVN, so we can then share it with others as well, because we have some projects where several people have to share the same bits of code.

I: Right, yeah. And are there situation where you have found a roadblock in the data, in the sense that you couldn't derive what you wanted? And how do you deal with that, with the researcher?

R: Well, we had a couple of instances where the researchers got a bit annoyed because we basically said the variable isn't good enough to use, and in the end we just need to find some kind of workaround. I think there have been studies that have basically been voided, saying 'No, we can't do this.' This happened, for example, with… we wanted to do stuff on child obesity, but we found out that the data we had was only the really early child measurements. When they have their immunisations, check-ups at one year, two years, and three years or so, but what we wanted were BMI measurements for adolescents, and they were basically not in this dataset, so we said, 'Well, unless we find another source for this data, this study cannot be done at the moment.'

I: Right, okay. Yeah. And so, this is more about the data you get from the GP practices, and what about instead when the researchers bring their own datasets? What kind of work do you do with them in that case, because they know their data and sort of…?

R: Normally the only thing we will do is project encrypt their data. Because what they normally want to do is link in with other data, which is the whole stronghold for saying it, yeah. So we get lots of questionnaires, the data's uploaded, Rohan's probably going to explain how that works, and then they create a project encrypted version of their own data, and then we can… and then they get an ALF and everything, so we can provide them with, then, another table that gives them, for example, hospital admissions for certain conditions, so they can be linked together. They will not see the original data in the SAIL gateway. It will also be in a project-encrypted view.

I: Right, yes. Of course. And so that means, in that case, it's much more straightforward, and you don't have to do a lot of these…

R: A difference in complexity of the projects. There might still be a lot of coding to be done to create the health tables for it.

I: Ah, okay. Like, for example…?

R: Well, we had the WOMBs project – Well-Off Mothers and Babies – they had a cohort of children where, in the early antenatal scans, you saw a shadow over the kidney, and at the moment, specialists don't really know what it means, but a lot of the women said later on the children had problems with kidneys, and they wanted to follow it up with our data. So what we had to do is create them, for their cohort, a bit table of hospital admissions for certain kinds of renal admissions. It was basically very long SQL script trying to re-code it in the categories they wanted.

I: Right. Because you couldn't… the original data didn't tell you… wasn't quoted as renal hospitalisation?

R: Well, the original data would be ICD-10 codes, and it's quite complicated. In hospital data, you get "episodes" and "spells". You can have several spells in one episode, so you actually start off with a person coming in, going to a different ward, being sent to a different specialist, and they're just all in the

same spell. What they only wanted is the first relevant code for each spell. I think I should have sent him a message saying I would be a little bit late.

I: How are we doing with time?

R: Oh, that's fine. I hope this phone is working in this office [laughs]. (Unclear 0.33.16) is working on the other side of campus.

I: So, yeah.

R: So yeah, basically some of the datasets are just very complicated, and they're awful. And first of all, the researchers normally wouldn't get access to the SAIL databank as it is, and they don't have the SQL skills to do all of the preparatory work to get it into a form that they want. So basically, in GP data, I think, there are over two billion rows of GP data, so we basically find the bits that are relevant to them, so they get a little bit of a smaller data set. But with the renal data, there was still one point three million records they got out of it, which was all admissions for the children in the cohort related to some renal problem.

I: Wow. Yeah. That was their own data?

R: But there was their own cohort, but the data we measured then against the hospital data, and of course, you get children with more than one admission as well, so you get… one child might had ten or twenty admissions if they have a serious renal problem, of they might start being on dialysis, or something like that, and they would all show up, somewhere in the data.

I: So, I guess that the length, also, of these kind of collaborations can vary?

R: Yeah.

I: A lot? How long is, and how frequent… how intense is the collaboration?

R: It depends a bit. We have some projects that we manage to do very quickly. We normally end up working with the same groups of key people, a lot, anyway. So we started with a card team for (unclear 0.34.56), and they're now doing, I think, probably about fifteen different projects on fifteen different topics with us.

I: Right.

R: But it can also be intensive. If there are problems with the data, there can be lots of communication trying to figure out the problems.

I: Yeah. [Pauses] So all these various views and scripts, is it easy to import, and learn from, the work of the other analysts? The colleagues… or is it more or less like… because there is so much specificity now, in what you do.

R: Yeah. It also depends very much on the analyst. Dan has a computer science background. He writes his scripts very differently from other analysts. A lot of us try to keep the scripts quite simple, while Dan normally tries to automate

as much as possible, and he uses SQL procedures and functions, that we probably should be using, but we were scared of them [laughs]. Not as experienced as he is in using them, and then it depends on how well people use comments in the script, or if they have an extra coding document that actually tells you what to do. So sometimes it can be very difficult to follow another analyst's thinking. We had a period when people said, 'Oh, can you just look over my script and tell me if you think this is correct?' But it's very hard to get into their head, and try to think out, 'Is this really the right way he approaches the data?'

I: Right. In that case, is that easier to do it to your own? Or maybe you go and talk, sort of…

R: It depends on the project. Sometimes your faster if you have the code to just sit down and start new. One advantage, as well, of that is that if you're looking at a lot of the same code at the same time, you're not freely seeing the errors. If you're trying to just say, 'That's it, I'm going with a new head on it. I'll start again', you've already learned from the previous experience, and you make it better, so you're improving your code. If you just use other people's code, and just use… so you're normally not improving the code over time, so try to improve them.

I: Right, yeah [overspeaking]

R: So what we've done with (unclear 0.37.29) as well, we haven't completely automated it, and the next step is supposed to be completely automated. What we've done is: Amrita and I started using our marked-down files, so you have the documentation and the running of the scripts linked together. I think you can do it in Python as well, but it would be quite good if ours… we did it that way. So you write a quite simple formatted thing. You can link in LaTeX code as well, to make the formatting a bit easier, and then you just have these little chunks of code. So say, this is the first bit of SQL script I'm going to run now, and if you re-run this one script, which does it all for you – it does the documentation, it runs all the tables, and it creates a cohort for you.

I: Okay. So you ran that script, and mainly in interactive documentation?

R: Yes, it's all very interactive. You have these little chunks, which actually you can link… connect to the database, send the command over to the database to create a table, and then come back with a result, so you can say, 'I'll do this now. The result is a table of eleven columns, and so many rows.' So we're trying to do that a little bit more. Dan has done amazing work with that. He did automated reports for GP practices, to give them some feedback, because they all signed up individually to support us and send in their data, and he's written these scripts which create for each GP practice their own report.

I: Right, okay.

R: So it looks like their data, it will show them what the overall statistics for Wales are, on how the individual practice is doing against this. And I think that's probably the way forward – trying to link it all together, because in that way you don't lose documentation, and if you have to do it again, you basically just press the button to re-run this one script. It's called "reproducible research", so it's a whole research field. It's been around since the nineteen

seventies, or something like that. So it's a guy call Knuth, I think. K – N – U – T – H. It's also "little programming", it was called, as well. So trying to link your documentation of how you do it with the doing it bit. And I find that bit really quite interesting. I try to do more of it. I've started doing… I also do some teaching, and I started using that for my teaching material as well, so I can give handouts where it looks like I've typed it all in, but I've actually run the code as well, as show them the results.

I: Amazing. Very nice, yeah.

R: So it cuts down on time.

I: Is this also… I understand now the technology might be behind those sort of data appliances?

R: I think data appliances… that's probably… a lot of that is still very much hard-cored with HTML and CSS mark-ups, so that's more of a programming thing. At the moment, we think it's something that's easier for the analysts who are not computer scientists to use.

I: Right. This… documentation [overspeaking]

R: Yeah, yeah. So that they aren't using [overspeaking]

I: Do you use a named release?

R: "R mark-down", I call it. "R" is a statistics [overspeaking]

I: R mark-down?

R: Yeah.

I: Okay, yeah. Okay.

R: So R is a freeware statistics package which is quite popular, and it links to all kinds of… people have done amazing things in all kinds of fields with it, but it just recently someone threw it all together in some freeware application called R Studio, and they've started having… one is Shiny. I don't know if you're going to say… Richard Fry would be a good person to talk to. He used Shiny, which makes great interactive internet pages. So you basically end up with an R code that does the bit of coding for you, and one script that tells you where the elements are on an internet page, and he used that to share information, spatial information, with other researchers. So, they had map data of, I think, traffic accidents or something, in Wales. But basically, because the areas are so variable – you have tiny little areas in the cities, and large areas – and the thing… he did an interactive map, on a web page, that he can share with other researchers, so you can zoom into the area to have a look at those tiny little bits that you don't see in a standard map.

I: Wow, okay. Very nice.

R: So I hope to do more with that technology.

I: Yeah. Amazing. And so, now, how do you divide between the projects that you have supported, in the analyst function, and the projects you're

more involved in? I assume you're not involved with all these projects at the same level.

R: It depends on the project. Sometimes I'll just have a supporting role, and one of the other analysts is actually doing the work. Amrita's doing the ACE project – Adverse Childhood Event project – at the moment, and she's been doing most of the SQL and documentation work, but she just comes to me to check on a couple of things. And other approaches are just [overspeaking]

I: So now you've got also a senior mentorship role.

R: Yeah, you could say that. We now have the… the kind of work we do, especially with the SQL, it takes quite a long time to train people to be able to work with those large data sets, so what we're trying to do is, as we get more and more involved in training the next generation coming in, and trying to explain… start off with an easy bit of a SQL, and then building it up.

I: Yeah. So, now, have you got programmes? Course and programmes? Do you train people?

R: Yeah. We have now a lecturer who runs a one-week course now on SQL, and we send the students there, and we've been having for a couple of years a visiting professor from the University of Western Australia, and he gives a course on health data linkage. Which is a very good course. A very tough one week course. So you have the introduction to data linkage, and then the advanced data linkage, and the advanced data linkage, it also goes more into methodology, and doing survival analysis, and actually analysing the data, and randomising the different cohorts. So a lot of the data analysts might never get to the point, because they're doing the linkage, so it depends if it's…

I: Okay. So is there a determined programme to train the new analysts, or is it more decided contextually?

R: Well, I think at the beginning, it was just up to us if a new member of staff had started on the same team, then the team would train them up. Now, we also just have the fallback, and just saying 'Go to this one-week SQL course, which is just in-house anyway, and at least do the introduction to data linkage course as well. But otherwise it's still up to the team. Some things might be very team-specific. Like we have a data-mining team at the moment – we don't have any training opportunities to tell them how to do it. So they just train their own team, and I think the mental health teams are saying they just train their own people up, to work with data.

I: Very fascinating. I would love to actually come to the course [laughs]. How often is it done?

R: Which one?

I: The weekly course in SQL.

R: Those are actually advertised on our webpage, there for external visitors as well.

I: Under SAIL?

R: Yeah. I think there's a [overspeaking] health informatics people, probably. Let me think where. I'm sure it's advertised through SAIL as well. They normally send an email round.

I: Okay.

R: We also got a new health data science course which only started last year. It's a master's course. That is part-time. My colleague Amrita's actually starting next week. It's her first week of study. She's all excited [laughs].

I: Yeah, amazing. Okay. I'll take that up.

R: The other course happens once a year, so I think it's probably in January. The introduction to health data linkage. It's a very interesting course.

I: Yeah, I'm sure, yeah. All very interesting. I'll check these out. Cool. I think it's probably… I've used enough of your time.

R: Oh, you're through your questions? Or do you have any…?

I: Yes. No, this is sort of a template, but it's very flexible, that I use, as a reminder of the challenges that I want to talk about, and the issues, and topics.

R: Yeah. If you have anything outstanding, you're round a couple of times anyway, so you can just pop in and ask a couple more questions if you want.

I: Another time, maybe.

R: [redacted]

I: Okay, thanks. Amazing. But I think we've talked of all the things that I wanted to talk about, which was what happens when the data comes in, what happens when you first go to a data set, and how do you work with colleagues, both in terms of the clinical researchers, but also in terms of the other analysts, what kind of processes for sharing learning and knowledge there are.

R: Well, one problem, I think, consists of one we have with statisticians is that they don't really quite understand the wealth of our data, and the problems with the data. So they only want… a lot of statistics are only allowed to do the statistics if the variable is filled in, which is normally not the case with administrative data, and the other problem they are finding is… I had this email for help, where he's saying… this woman trying to do analysis in Stata, but programmed since… if you end up with two million rows of data, then certain statistics packages can't deal with it.

I: Okay, yeah. Yeah.

R: So it gets very complex. You're trying to link another data set in, and then the computer freezes over, so that is probably one of our largest problems, if you want to do very, very large statistical analysis, but the software packages themselves can't deal with it.

I: **Yeah, and you need to scale up your…**

R: Or you might have to completely write it in a parallel progressing language like Python, where you can actually send chunks through, or use a cloud behind. I think they're setting up a Hadoop server at the moment, so we're hoping that we can use Hadoop to do a lot of research which is on a normal computer, just not powerful enough.

I: **Yeah.**

R: I think the worst are our data-mining team, because they look at everything in the GP data, and compared then five hundred variables with another, and they have problems with running out of memory, running out of space, having to re-write their script to adapt for this kind of thing.

I: **Yeah. What is this team, you said?**

R: I would say Shangming Zhou is the head of the team, and then you've got Fabiola Fernández-Gutiérrez, she's one of the analysts, and then John Kennedy, who was phoning, actually, but he might have gone to lunch.

I: **Okay. And they work with GP data, is that?**

R: Not only GP data – also with hospital data, but they basically look at as much as they can to try to understand the coding of arthritis. What codes are you likely… if you end up having a specialist meeting for arthritis, what kind of codes do you tend to have before you go there? So they use (unclear 0.49.58 – 0.50.00) black box approach for it. And that is very, very memory intensive.

I: **Black-boxed.**

R: Yeah. Very black-boxed, but also very… I don't fully understand the data mining. What they end up with is like a training dataset and an evaluation dataset, so they can compare the outcomes of the model.

I: **Yeah. Very interesting. Thanks. I'll actually leave you, and try to talk to them as well.**

R: Okay.

I: **Thanks a lot for your time.**

(End of recording)