NTDS 014

Key:

I: Interviewer
R: Respondent

I: Okay, thanks for your time. As I told you, I'm interested in… have a sort of, first, a general understanding of the role of the analyst in SAIL, then you could maybe introduce yourself and your background.

R: Yes, sure.

I: - and then as a sort of trajectory to how you go involved into SAIL.

R: Yes. [redacted] I'm a Research Analyst at Swansea University, working with the SAIL data bank. [redacted] I really enjoy the software development/data analysis side of those subjects. I was there just applying for jobs while I was finishing off my Master's, [redacted] and we were building the new ILS building, and this SAIL project was going to be quite big, so I looked on jobs.ac.uk and I found my role advertised. And my role in particular has changed a bit but it was working [redacted] in, essentially, epidemiology-based studies. Each of those two groups, they had clinicians and academic staff directing research but they required a data analyst to essentially use the SAIL data bank, pull data together that could be used for… produce datasets for analysis, and that boiled down to initially just coming in each day and writing SQL queries. But then my role has kind of evolved a bit since now. Instead of just pulling together datasets from… the core datasets in SAIL, I get involved in writing papers and deemed first author, so that entails doing the stats analysis, participating in study design, and doing literature reviews and all that stuff.

I: Okay.

R: So really I go right from the beginning, where I'm using SQL to isolate bits of data, all the way up to writing first author papers.

I: And these are still within [redacted]?

R: [redacted]

I: [redacted]

R: [redacted]

I: [redacted] **So the job advertised, are you involved in SAIL full-time? So the job is retail, or is it…?**

R: So my role is a bit strange. Now, half of my role, I am a core SAIL analyst, I'm employed by SAIL, but before that my job was… it was funded by the medical research councils and charities but I was deployed in SAIL. But now [redacted] I'm

now a core analyst in SAIL. So I'm supporting other people's projects as well as doing my own research that comes our way.

I: Okay. So you are a core analyst. Okay, so the writing papers, study analysis and literature of these sort of functions, these are sort of part of the service that you do as a core analyst per se and there is part of the other bit? I'm trying to sort of understand, so.

R: I don't think it's possible to put them in two different bins. Even in our group meeting yesterday we were discussing this exact question – are we just going to support projects coming into SAIL or are we going to do our own research? And really the answer is: we are going to be doing bits of both. And because everyone, even within the core SAIL analysts, there's a team of, I think, six of us, six or seven of us. We all have different interests and we are allowed to pursue those. If you are more research orientated than you can produce academic papers. If you just want to do support and clean databases and write bits of software to facilitate that then you can do that. I enjoy doing more research, but I do, you know, helping, making sure research is reproducible and making sure everyone is using SPN properly, version control and things like that. So I do a bit of everything really.

I: Mmm. So you offer then also training and these kind of things, if you make sure that everybody is doing SPN stuff? Do you offer support and training?

R: Yes. So some are more official than others. [redacted]

I: Okay.

R: [redacted]

I: No.

R: [redacted] in data mining. So obviously that way, that's official support, and we run seminars, weekly seminars, just the seminar using our programme language and a data mining seminar. But a lot of it is quite formal. This research group is doing this project and they want to know some best practices and then we can help. But usually they come to us; we don't go around [overspeaking] people.

I: Okay. But also then you collaborate also with the students, you involve also the students in some of this work?

R: Yes, so actually I have been involved with helping some students, really medical students, in their third year, when they are doing their third year projects, and those projects that they have done have turned into papers as well.

I: Oh wow.

R: So they all come in with an idea from their supervisor. One was a PhD, actually. I'll just… I can either just do the data analysis for them or I can help them with the study design. So I quite enjoy that as well.

I: Great. So your interests have also been a little bit changing in broadening, right, because you said you came from the computer science, this side of it, and now you are more interested also in health research aspects.

R: Yes. Well I just knew that [redacted] I wasn't really interested in academic theory and in data as such but rather applying the programming skills towards data, and health informatics is just a brilliant way into that. [redacted]

I: Yes. [redacted]

R: It is. In a way it is. My day job is health informatics [redacted] But actually some of the studies, they can be linked together. [redacted]

I: Okay. So it's still about linking?

R: Yes.

I: Interesting. So if you tell me a little bit more about… in more detail about your pipeline. How do you get involved in a project, do people contact you directly or do you get sort of… is there an (unclear 0:09:08.5) by some SAIL staff, and then how does it work out?

R: It's all sorts. Like in my role, I think I'm quite lucky, I've got a lot of freedom because I can sort of speak with colleagues. You know, you meet at a conference, for example, and you meet someone with similar interests and you can just get in contact. And the problem is then how do you fund potential projects? I've obviously got my bosses, which they say, 'Well this is the project we are working on.' But today, for example, I'm meeting with a clinician [redacted] he's put me in contact with his boss, and I've proposed a project to get this data in and there's going to be a few pilot studies on that patient that has been followed-up from the GP from the hospital prescriptions. So I can actively seek out research providing I can find some funding or at least get grant applications. Or the other side of it is, especially now, some half core analyst projects, I'm sort of delegating projects from what we are funding to do.

I: [Overspeaking]

R: Yes. So there's that side as well.

I: So how does the collaboration with the users work out? Also, because I can expect… you probably have worked on a range of different projects, and I think what can be interesting maybe is that you come from a very

deep expertise of software methods and probably your user has got the other side of…

R: How do you define "user" in a sense? Because I guess I would view myself as a user of the data, so.

I: Okay. So you see yourself as a user, definitely, but more the health researcher that is approaching there from externally.

R: Yes. So we have clinicians [redacted] that will say, 'We are interested in this project.' It usually involves just a conversation with them explaining what SAIL is and is like a feasibility conversation then – 'Can we carry out your project?' And sometimes that will involve the health researcher providing their own dataset, which they are interested in, which can obviously link into SAIL, and then once they are happy to do that, they will generally have a broad research question because they want to answer, and then we'll work together to chisel it down to something which we can definitely deliver. I'll handle the data analysis even at a… So I'll do the data linkage, the analysis, write parts of the methods and the results up, and you are collaborating with these health researchers the whole time.

I: Yes. Definitely. So what are the things that you need to ask, and do you need to ask collaborators in order to do the feasibility study? Because that's, for example, an area that I want to understand better, how it works, that process.

R: Well, the first thing we normally ask is: is it a Welsh dataset? Because some hospitals we have in Cardiff, they actually have people from Bristol coming in. So it's a geographical question that you have to ask – 'What's your demographic?' And also what…

I: Why is that important?

R: Well, so SAIL, for example, they have GP data and hospital data but generally it's Welsh data that most of the analysts work on, so we'd have to check is their cohort compatible with SAIL's?

I: Otherwise you can't link it?

R: Exactly, yes. So there's that issue. But some people don't provide their own datasets and they are more interested in stroke and diabetes, for example. You want to ask them, 'Well what service or access to care do they have?' 'Are you looking at GP level, hospital level, are you looking at prescriptions?' I don't actually have to ask a lot of questions, the clinicians, they just come to us and they'll say, 'I'm interested in this,' and then, 'Oh we can do this for you, we can do this for you,' and latch onto that.

I: Okay. So what are they usually looking for then into SAIL? Do they have a clear understanding of what they can do with SAIL? It's a pretty complex system.

R: I think, yes, well in recent years a lot of clinicians, especially around ABMU health board and Swansea/Cardiff, they are aware of SAIL and what it can offer. They know that in their disease area they can get a nice epidemiology

study out and they'll approach that, or they will say, 'Well we are interested in this, can we write a grant?' and facilitate the research that way.

I: So do you have lots of people also that contact and want to work with SAIL from outside Wales or from pretty far in these terms?

R: Yes. So we get pharmaceutical companies - I'm not sure I'll be naming them here - and within Europe. And some of those come through a collaboration with Welsh government. So yes, there's definitely outside of Wales interest in SAIL.

I: So, do these projects get carried out, for example, in a feasibility… because of this sort of situation of finding out, so data requirements and sort of the…?

R: Yes. So [redacted] a scoping feasibility exercise. That was a pharmaceutical company. They just came down with the representatives and they seemed to know about SAIL already and they just proposed their project which involved around their products. Yes, it's just sort of you get round the table and… Then we have scoping forms where you can send to researchers and say, 'Well, what data do you want from SAIL?' and then what we can offer. So you can do it more officially if you want, but usually you start off with just conversation round a table and then you start filling in all your scoping forms from there.

I: Alright. So is there much interaction? There's not very much interaction between you and them about what they are looking for and what do we have and [overspeaking]?

R: I mean, this is only from my own experience. [redacted] Clinicians are different because they are local and the lecturers are around here, and you can just pop over and see them now and again, so you do… Sometimes you don't get it all in one meeting. So I'm not too involved with scoping these research interests outside of the UK. [redacted]

I: Okay. Yes. So then these projects can sort of start in quite an informal way, I guess, and they don't have a very specific timeframe, I understand, because you say I can meet them several times.

R: Yes. It definitely really is informal. Even things like sitting next to someone at a conference, that would be a good idea, send a few emails and…

I: Yes. So what kind of requirements have you been unable to satisfy sometimes, if at all?

R: Um…

I: Because I guess that you get an understanding around some queries to find out sort of what you have there.

R: I think the only thing that springs to mind really is lack of data. You get lack of linkage, but also there's been a few times where some researchers or health researchers that I've worked with are interested in a disease or a drug and then you look at, well, how prevalent is this disease within SAIL data or how

often is this drug prescribed in SAIL data? You just don't have enough there really to have enough power in your statistical analysis. So yes, it's just data linkage, and there's just not enough data there really.

I: So how do you define lack of linkage?

R: So I guess if someone has their own cohort that they are interested in, particularly if they've looked at a list of patients in their department in hospital and then they will come to SAIL and they'll want to link in to all the other datasets that we provide. I've never actually had it where a study has not been feasible, but you get that cohort sizes may reduce by 40%, 50%, and you always have to explain why that is the case, but usually it's fine. I've never actually had a project just stop dead because of that reason but there's potential there to happen.

I: Right. So the cohort size will maybe reduce because of prevalence of this interest, disease and travel interest and other reasons.

R: Well, there's a geographical thing, because, for example, if you want to look at prescriptions, you have to look at GP data currently in SAIL, and so we have… it's about 70% of Welsh practices signed up to submit their data to SAIL. So if you've just got a cohort of people in certain hospitals, for example, Cardiff, we don't have as good a coverage as Swansea, so if there's a department in Cardiff, then more of their cohort will be missing when you link it into SAIL. So there's that. But then, yes, so if you look at a certain drug that only came out two years ago, you are not going to get much.

I: Right. Yes. I understand. So how much technical familiarity do the people that work with you have with the datasets -

R: Yes, again it varies.

I: - as opposed to the problem domain, [redacted]

R: Yes, it does vary. [redacted] he's got a background in software development and a few of the public health researchers that we work with, they are always really good with data anyway. They can even write your own code and do the data linkage themselves. And then we offer a course which… health researchers just can come to us and we'd offer that course to teach them about data linkage, but then you get sort of professors and clinicians particularly that are active in hospitals and just don't have the time to invest themselves in knowing about… they know that they can put their faith in other health researchers just to inform them of what's feasible in terms of data, but they won't know how to do any of the technical stuff. So yes, it really varies.

I: Right. So how does it work out when the people don't know how to do… or the technical staff? Is it harder to understand, or are they having a harder time to understand what the issue is here (unclear 0:22:04.8)?

R: Personally, I've never got to that point. I think maybe I just explain it well, I don't know. Yes, I've never got to that point where a researcher has just not really understood how we can help them.

I: Yes. Do you then carry out all the analysis also to the SPSS or the stat (unclear 0:22:26.9) packages that are used?

R: Yes. It varies what project. I'm happy to do… I'm not a statistician, but I can do some stats. We have some statisticians in SAIL, you can just speak to them and make sure what you are doing is sound. Then some health researchers, they have that background anyway.

I: So, and what is this course that you offer to…?

R: It's just the beginners and advanced data linkage course which we have been providing, and we collaborate with University Western Australia, with D'Arcy Holman and David Preen. So they've come over for the last, I think it must be five years, and they offer a really comprehensive course on data linkage and that's actually quite popular with health researchers, we've found.

I: Yes. So that's a course at the university level, or is that for the SAIL users?

R: It's a SAIL course; however, we've now incorporated parts of that course into the MSc Health Informatics.

I: Okay. And so it's a course that you offer then also to the researchers to collaborate with SAIL?

R: Yes. I mean, anyone can…

I: That's the target? This type of course is for the people that…

R: Well it can be used to provide supplement training to employees that will be working as an analyst for SAIL, but then, yes, it's also for health researchers that want to gain some knowledge or even do the research themselves.

I: Okay. But it is primarily for the employees, it's a training course for employees for SAIL?

R: Yes.

I: So how many people are sort of, you know?

R: Per course, I think you are looking sort of 25, 30 people now. It's quite popular now.

I: And it runs every year?

R: So, to my knowledge now, the beginner's course and the advanced course is one, so it's two courses per year.

I: Right.

R: One beginner, one advanced.

I: Oh wow. I would like to see some of the material, it would be great.

R: Yes, no, it's good.

I: So then in terms of the infrastructure, you are able to do everything from your own computer? You need to sort of be tied into some infrastructure?

R: Yes. I'm sure the SAIL gateway has probably been mentioned to you. So that virtual use of the gateway is just virtual interfacing. It's got all your tools there. DB2, SQL, Explorer. So yes, really you can do all your analysis just from there.

I: But you (unclear 0:25:36.3) or not?

R: No. We abide by data protection laws, but we are not… me as a researcher, I'm not personally a guardian of the data, I just have to work within the data protection laws. SAIL has its own agreements, it's own data agreements. When we work with clinician in certain areas, that whole data, you have Caldicott Guardians who are responsible for okaying us to use their data. Yes, I guess I'm not personally a data guardian in that respect.

I: Okay. So you will ask somebody else to… for the exportation of it and things?

R: Yes.

I: Because, for example, what I'm… I'm not explaining myself good enough, probably, but I read the paper on the gateway. There's the diagram of the structure of the gateway, and the analysis can be carried out by things like (ph: 0:26:43.5) PPPN and also some data can be exported. Because you as an analyst, but then also a researcher and as SAIL staff, I'm trying to understand that you sort of… if you are… that's why I asked what kind of role do you cover.

R: So its best practices are research. We receive training on how to avoid disclosure of data and we have another layer so that any data which come out of SAIL or has had to be reviewed by a panel, we have information of governance panels where if you are interested in using a certain dataset, you have to provide your analysis plan and that will go off to the IGRP panel, which is an independent panel, and then they will okay that project, or if they have any concerns to raise about possible identification, they'll bring that up.

I: So you do not produce the data views. You use data.

R: I use data views, yes, but then we refine those data views into a data which can be used… you know, you can run your starter or SPSS with it, but we don't export that data out of the gateway, it all stays within… So any results which come out, they are, as you would read in a paper, you've got your statistical analysis. There's no [overspeaking] data. So in that sense, yes. I don't know. We work within strict guidelines. We are not a data guardian.

I: Yes. No, no, no. What do you mean by when you say we refine the data views? What kind of requirements are you working towards to refine it, the rationale?

R: So yes, you have data views provided. So for example, the GP data we hold in SAIL, that's presented to us as views, but then you want to look at a certain disease and a certain timeframe or certain locations and so you use an SQR then to pick up any records relevant to your study, and then if you get into a nice clean dataset, it will just pass through SPSS. So I've created that. It's a view, in my view.

I: Okay. Perfect. Okay. So you strip out some of the unnecessary bits.

R: Well, the research question, you've got this huge dataset with all sorts of diseases and you are only interested in one or one drug and so, yes, of course you have to strip out on the record you are interested in.

I: Yes. That's part of, I guess… I understand now that that's maybe part of the core of the things that the analyst does, right –

R: Oh yes, definitely.

I: - for the researcher, is like there's a bit of you that you are accepting it's more complex and there's more stuff and then you, according to the conversation you have with them, sort of you just (unclear 0:29:53.8).

R: Yes. And sometimes a lot of it is you have to join data views together, core data views within SAIL and then the view you create is then representative of that join.

I: Okay, yes.

R: So we do produce datasets like that.

I: Right, okay. So because the core data views are sort of (unclear 0:30:19.5) data, it was important that they are separate for the various databases, in that sense?

R: Sorry, what do you mean?

I: You said you are joining different tables from various datasets [overspeaking]

R: Yes. Assuming you want to look at GP care and then how often people end up in hospital, you have to link those views together. So yes, you've got your building blocks, your core datasets and you just take the bits that you want, join the bits together and produce your research dataset.

I: Right. And so, after the data research study, what do you do with that view, will you save it, is it that reusable?

R: So yes, you want to keep your data views that are available if you want to go back in two years' time and query that dataset. If you want to say, 'Does this still stand?' then it needs to be there available. It's like reading on your form that… on your datasets that they are destroyed after a certain amount of time. So I think… I'm actually not sure how many years you would keep it then, but quite possibly, off my head, maybe five years you keep the dataset there.

I: Yes.

R: But then you've also got the code which produced the dataset which is probably more valuable.

I: Right, yes. So, and is that shared across the analysts? And you can also take out jobs from, you know, continue jobs that somebody else has done?

R: So some pieces of code or software, some projects are restricted to only the analysts that are on the project, but then we also have, between the core team, we develop bits of code which are… they are tested, their best practice codes, which you can use for your own project -

I: Mmm.

R: - and so that goes hand in hand with trying to make all our research reproducible.

I: Yes, yes. So you've got it sort of the library of…

R: Yes, the code, in a sense, is more valuable than the dataset.

I: Yes. No, of course. So then you review it across the core team members?

R: Yes. So in our team, the core team, we review it, we do unit testing, cross validation, and once we are happy with a bit of code then we are happy to say, 'Right, you can use this.'

I: Okay, yes. So that's sort of for new stuff and then that's also for adopting things as part of the library?

R: So by that do you mean… if something has already been done, someone has already done… answered a question, which you and I want to borrow that…

I: Yes, how does all the knowledge that's embedded in a script get really shared? How do I know if a core analyst, what you have done?

R: Yes, we have a wiki that we administer, which will contain all bits of code which we are happy that have been tested or that bit of code has been used in a paper that's been published, and that's an internal wiki, it's got a lot of stuff. It teaches you how to use different statistical software.

I: Right. And so every one of you sort of publishes the stuff on a wiki and then the others sort of review it?

R: Yes. You can put your… we would present a lot of our work at our data analyst meetings and you are putting your code out there. And I can't speak for other researchers, but when I'm doing research… within the team I work in, we would share our script, we would test it and make sure that you come to the same answer, write the same results, and then when we are happy we'll then release that onto the wiki so people can access it.

I: Yes. So more than one person tests it. So this wiki, is this the same wiki that is the wiki of the gateway, or just different wikis, because I haven't seen them yet?

R: Yes, it's in the gateway, so it's in internal, because the code you put in could result in sensitive information being released. So it's all about the wiki's internal…

I: Yes. So it's another one, yes. And then there's one instead for a gateway which is also for the other people that are the users.

R: No, so we only have an internal. So we've got a data applied outside the gateway which will provide to health researchers what datasets SAIL has and a list of published papers that have come out of SAIL. But in terms of…

I: In the paper about this, the gateway, there was a description of sort of data analyst suite that is part of the actual desktop and it was… part of it, it was talking about a wiki.

R: Yes. So we have the wiki, which is inside the gateway, but it's not external. But we do have a wiki, and I guess that's the wiki that you are talking about.

I: Right. Yes, okay, so it's the same. So these data analyst meetings, how often do they happen? What do you discuss at these meetings?

R: There's loads of different meetings depending what group you are in. We have monthly SAIL user forums. Quarterly we have… quarterly SAIL user forums but external health researchers are invited and then I guess you have your project meetings. So the core team, we have meetings every fortnight, we have code reviews at 9:30am on Monday, so we go through each other's code.

I: Okay. That's for the SAIL analysts?

R: Yes.

I: A-huh. And the project is for the fortnight. Okay. When you produce these views, sort of refine views, do you produce also new metadata, new descriptions of derived data as well for your users, or you simply sort of copy and paste select columns from tables?

R: Yes. On the whole you are just producing columns which exist. I do a bit of natural language process in work, but it's really early, but in that I can definitely foresee that we are creating our own data. But on the whole, as a core, my data analyst work that I do day-to-day is not really deriving new variables.

I: Mmm.

R: I'm trying to think. I suppose maybe one example I could give, I derived a cardiovascular risk score, but I was only following instructions… so it's QRisk. So QRisk published their guidelines and I replicated that risk calculator using SAIL data. So I guess the final score is derived in a sense but not something that I've derived on my own.

I: Yes. It's a sort of comparison with [overspeaking].

R: But then I included what we call a data dictionary – this is my table and… And actually, the data dictionaries are helpful because even though analysts understand what columns exist within core tables, external health researchers, they won't know, and so usually when I provide a table they provide like a list of this column is this, like write it in words and so on.

I: Yes.

R: It's not necessarily deriving.

I: So you produce also this sort of documentation.

R: Yes.

I: Oh, I also wanted to ask you, so in each project there usually is one analyst or does he (unclear 0:39:35.7) you need to work with other analysts, you need to involve them depending on kind of the problem that you are solving?

R: Yes.

I: Is it one analyst per project and that's how the SAIL analysts [overspeaking]?

R: Sometimes it can be one. In the core team I'm in now we usually shadow each other, so we try and get two, even more. On one project you can shadow each other and unit test each other's code and so on, but sometimes a funding can only afford one analyst, in which case it has to be only one, but then it's up to you then to present your results to clinicians, and they all say, "Well that doesn't quite make sense." There's always a sort of back investigating what you've done and then comparing to literature out there, does your result match, or is it in the ballpark of what's already found, and extensive literature reviews making sure you liaise, you are in the ballpark of what your research is.

I: Right, yes. Can you elaborate more on this?

R: Yes. So particularly a lot of these projects revolve around a certain disease, so you need to make sure that… well, first, before we do any analysis, are we picking out all the records we need for a certain disease? So you have national figures, and all literature will say 'The prevalence of this disease is that…' Because in certain datasets we've got all of Wales, like hospital admissions, you can check your result against national audits and you know that you are in the ballpark, you know that it actually matches really well, and within a couple of people out of tens or hundreds of thousands, so you know you've got the right answers. So there's a lot of that going on with it.

[redacted] So in a GP setting, GPs will use read codes. So we use those exact codes to pick out [redacted], so we are always constantly comparing our answer to what's already in the literature to

making sure we've got the prevalence, otherwise if you are short you know you are not picking up all your records.

I: Mmm.

R: So yes, there's different ways of validating. It's obviously great to have two analysts who share each other's code, but that's not all the validation required, you often have to look at the literature to make sure you are arriving more or less at the right answer.

I: Right. The right sort of initial commission.

R: Exactly, yes.

I: Right. Okay. So in that sense, that's also when it's more important, if it is a new topic for you to work on and sort of communicate with the…

R: Yes, especially in rare diseases. And often there's not a lot out there, you just spend a lot… When it's a case of there's not a lot of literature out, you spend a lot of time consulting with top clinicians in that area and get them involved.

I: Right, yes. And so what kind of feedback have users, your own users, sort of the clinicians, sort of given to the SAIL process and what would they wish to be, you know, more developed towards?

R: I think, obviously, to actually do the research, day-to-day research in SAIL requires a certain technical expertise. I get the feeling maybe some clinicians would like that they would have some tool, a graphical tool where they can query the data themselves. All I've really heard is positive stuff. I'm trying to think of ways clinicians might want more out of these sort of platforms. There's always the push for more data, different datasets, richer datasets.

I: And also, maybe so in terms of administration, costing or… I know that there's been… SAIL has learned how to cost an analyst over time, no?

R: Yes. So we use a timekeeping features. Have you ever heard of Toggl?

I: No.

R: You can actually just record… there's an app on your phone and you can record – right, I'm working on this project now, and you can track what projects you are working on throughout the day so that when a company, a pharmaceutical company or a researcher comes towards you and says, 'Well, how long will it take to do this?' we can look back at similar projects and we can say that took this amount of person time to do that. So yes, there is that, and we can provide that to inquisitive researchers.

I: Okay, yes.

R: I guess a lot of health researchers also want… if you imagine you are a GP in your practice, we've started releasing reports which compare their practice to the average in Wales, so we are actively trying to seek out what health researchers want other than academic papers, which is one of the thrusts of SAIL. Yes, but we have a lot of these surveys. [redacted]

got together clinicians, carers, patients [redacted] asked them what research topics or what did they want to know more about of their disease. So a lot of that feeds into SAIL and what priority of research areas we have.

I: Okay. So did that give the input for some of the projects in the end?

R: [redacted] yes, that was… in fact, all of our initial questions are based off that and they've sort of evolved since, from that one survey.

I: Okay. Mmm. And in terms of data importation, when new databases get integrated, have you done also work in that respect?

R: I don't do the technical way to do it. I get involved in speaking to clinicians that have datasets and explaining what they could do if they linked it in with SAIL. Generally, that's more of push process on their side, so they can push the data over to us and we facilitate that process and give them the guidelines.

I: So you explain the process, what they need to do.

R: And what happens to their data. There's like a diagram in the paper, you get the split-fire process. We explain all that to them and how their data is going to look on the other side and what they can and can't do. So we explain you can't just backport your data out of the SAIL gateway. Once it's linked to other datasets you can't just take it all out. We explain all that. So I've been involved in speaking to clinicians and that side, but then obviously the technical team at NWIS, they will process that data. I don't see that.

I: Right. But then do you review the data that has come in -

R: Yes, Q & A.

I: - for quality, yes?

R: Yes. I've QA'd a few datasets. [redacted] and that will normally involve one or two people and QA'ing that. Producing a document, this is the dataset.

I: Okay. What's in the document?

R: So explaining these are the columns, each column write down what the variable is and what coverage it is, what timeframe it is.

I: Right. And so that's a description of the dataset after it's been put into SAIL not as it was coming in?

R: Yes, afterwards. So once it's all anonymised. And then what's provided for, SAIL users to use… Of course, some datasets, they are restricted to only certain projects and some can be used… That decision is made by the data provider, whether they want their data…

I: Right. And do you give also an evaluation on this data, that you say, 'Okay, this column, it's got diagnosis codes'? Do you still make observations about [overspeaking]?

R: Yes, you can do a quick analysis, and there's a lot of starter packages. You can do summaries of each column, like averages and medians and things like that, you can look at how many women are in this dataset and drug prescription data. So yes. And really that involves just going through every column and explaining what it is and what kind of coverage that variable has, you know, with this column it has a lot of missing data, for example, and it's not reliable.

I: Yes, so you explain that.

R: Yes.

I: So you also review the data for inconsistencies and sort of missing data.

R: Like typos or the date is in a wrong format. So you do sanity checks.

I: Right. And how do you do that, do you have automated tests, or do you have a look?

R: I just have a look. I don't have any automated tests because each dataset just can be so different and automated tests might not pick up… you know, too much relies on automated tests. You can build… Even though you are getting certain columns in, you can build in some automated tests, but usually I just look at a dataset, I'll write my own notes on it and document. I'll present it back to the clinician or the data provider. And often some columns I don't understand what their abbreviations… I don't understand what they are, but it involves me just looking through it.

I: Yes.

R: I think some columns of data can be impossible to embed in an automated test. It won't catch all, so.

I: Like, for example, what kind of data is difficult to…?

R: Well, so if you've [redacted] results, I don't know whether you are allowed to have minus results. They always seem to be positive results. I don't know how large that scale can be, you just have to look at it and go, well this summary, the largest value is this, the smallest value is this, go back to the clinicians, 'Does this make sense?' So there's a lot of communication with the data provider to make sure that the data is okay.

I: Okay. So what is the communication, how do you sort of…?

R: Email, phone…

I: So you ask for explanations?

R: Like I said, if it was [redacted] test, that was just an acronym or an abbreviation in there, the column header, and I just didn't know what it was. So yes, it's a lot of emailing back and forth mainly.

I: Right. And you send them the documents as well so that they can comment and suggest?

R: Yes. So the document would be held within SAIL. Personally, all the health researchers I've worked with, or at least one of the health researchers for each project, it can be a few projects, at least one of them actually has their own SAIL gateway account because they are interested in learning more about how you analyse it. So for me it's handy, you can just see the documents there. Or if not, I just have conversations on the phone and just say, 'Well, you've got three agendas in this column, what's this about?' and things like that.

I: Amazing. I think that's all really because we've covered really many things in a bit of time, yes. Do you need to… are you pressed?

R: No, I'm fine.

I: It was very interesting. How long are you now in your position?

R: [redacted]

I: Yes. And what's the plan after, you are going to stay?

R: Well, I'm really interested in…

(End of recording)