161020-150801

Key:

I: Interviewer
R: Respondent

I: Okay. Recording. Yeah. Okay, so would you like basically to just introduce the project, how it… what it's designed to do, how did it come about and if there was anything distinctive about it or… yeah.

R: [redacted]

I: [redacted]

R: [redacted]

I: [redacted]

R: [redacted]

I: [redacted]

R: [redacted]

I: [redacted]

R: Okay.

I: Yeah.

R: That's cool. So basically the idea was to really [redacted] use of the resource [redacted] in SAIL, the big administrative linked databases to help records and everything, [redacted] this big, black box system that does this anonymised data linkage that is really significant for the whole of Wales and then to see what we could do in terms of how can we link the data [redacted] to the data in MEDMI, particularly around weather and environment and so on.

I: Yeah.

R: So the project that we came up with was all around childhood obesity [redacted] So we had data on childhood obesity from the Wales Childhood Measurement Programme, so school, primary school measurements. And then the idea was basically to link that to green space data [redacted] and weather data from MEDMI with the idea that there might be some sort of interaction between whether environment in the neighbour and kids' weights data, so basically

there was quite a lot of work to be done in terms of… [redacted] in terms of extracting the right data and as with all these things… I mean, you were just talking about in terms of the spatial data was the critical bit of it really and actually to be able to pull out data from SAIL at the right geographical level was actually useful and it could then be linked with data from MEDMI was a big job for them.

I: Okay.

R: Basically, what we ended up with was some linked data, which we were able to analyse, but we lost so much data through [redacted] need to suppress data and with small area geographies of Wales. That… we felt that it was just massively biased, so we didn't go… we didn't really find any interesting relationships -

I: Right.

R: - in terms of what we hypothesised, but our conclusions were that we couldn't really draw any conclusions from that because so much data had been suppressed and it was likely that because of this… the nature of it, the data that had been suppressed, were areas where you primarily had very low rates of obesity and consequently smaller numbers, so a highly biased -

I: Right.

R: - suppression of the data, which then meant, yeah, we couldn't really do anything with it.

I: Was that particularly important for the analysis because these areas with the lower populations were also the areas where you would have been more likely to detect the effects that you wanted… the relationships that you were targeting?

R: It's hard to say exactly. I think that's part of the problem that you couldn't really disentangle it and say where the bias is likely to come from, but it's likely to be… yeah, like you say, the fact that data are more likely to be suppressed where you've got small populations and perhaps those areas tend to maybe be more rural rather than urban… data also tends to be suppressed where you've got small numbers of kids who are overweight and so they are actually… you know, that's the critical one really because if our main outcome measure is what percentage of kids are overweight, we've lost the whole bottom half of our distribution, but we can't really draw any conclusions about relationships with the prevalence of overweight. Yeah, so it presents various challenges, but that's the key one.

I: Yeah. So this key challenge was really related to the fact that you were trying to detect and understand something that has got a not very big prevalence in the population, so it… the obesity is the minority in a way.

R: Not really because it's still... we're looking at reasonably high prevalence on average, I can't remember what it is.

I: Okay.

R: Twenty, thirty, forty percent, but even with that, in some areas where you've got low prevalence, so where… effectively where the kids are healthier, the data gets suppressed because there's not enough overweight kids.

I: Right.

R: So it's actually a function of these areas being healthier that means their data gets suppressed, and obviously we're trying to look at the geography of what makes some areas healthier than others and so it becomes a sort of impossible question to answer if those are the data that are being suppressed. We randomly lost… some are done. Losses were forty, fifty percent of areas. If we randomly lost half of our areas, it would affect it, but we might still be able to see some relationships, but because we're not randomly losing them. They're totally biased what we're losing, it affects our ability to draw conclusions. But we had… this is also partly a function of the way we were going about it and the amount of resources and time that we had to do it. In an ideal world... and the idea for potential from that work would be to take data from MEDMI, import it into the SAIL databank and do the linkage within the SAIL databank in the anonymised setting -

I: Yeah.

R: - where therefore, there wouldn't be any suppression involved and that could… that would be the solution to using this data in a way that actually worked.

I: Yeah. So in a way, that makes a case for this kind of secure infrastructure like SAIL where it's easy to take in stuff and hard to take it out to become aggregators of those kinds of research.

R: Yeah, absolutely, but that all requires the relevant data owners to be willing to allow their data to leave their system and be incorporated into somebody else's system.

I: Yeah. Yeah. I guess the part of an issue was also… I think, probably, there was an interest to try the MEDMI data in such as they are in MEDMI because it was a pilot project of the MEDMI.

R: Yeah. Yeah.

I: Yeah. Yeah. Yeah. Yeah. And so you were saying in terms of... you were saying that there was quite a lot of work in extracting this spatial data. It was critical in order to get this data to be useful. Would you be able to expand more on this? What it meant. Maybe did it make the project run slower? Did you have to take decisions about how to… what granularity to work on…

R: I've got to say that [redacted] I'm not really familiar with the exact process. I think it was fairly standard. It was fairly typical [redacted]. But it required one of [redacted] analysts to do this pre-processing of the data and make… getting it into a state where… going through their procedures to allow it to come out of SAIL and just producing the actual data that we needed, so I don't really know much of the detail about it, but that's where the bulk of the resource that we had for the project… most of it went to actually funding that.

I: Yeah. Yeah. And in terms of interpolation between these three different types of data, was it fairly standard work or did you have to take… how did you discuss what level to work on? Comparing datasets.

R: I can't remember how we exactly made the decisions. I think partly because we wanted to… as is usual in these things, we probably wanted to make sure that we could link the health and environmental data to socioeconomic depravation data, so that set our geography as lower layer super output areas. That's what we wanted. And yeah, that was it really. That was most of the decision made for us.

I: Yeah.

R: [redacted]

I: Yeah. So the way I'm trying to make sense of this is because there's such an important case for deprivation data, right, because it's a very strong factor, parameter, to influence the obesity.

R: Yeah.

I: Then the choice for lower super output area is done despite it is... it imposes then a minimum common denominator in terms of resolution of the location aspects. Is that correct, thinking of it in this way?

R: Yeah, that is true and it is definitely part… it is definitely the main driver, but it's also… I think because especially with this approach, the smaller you make the area, the more likely it is the data are going to get suppressed because you've got smaller… a smaller number of people basically, so we could theoretically have done this piece of work at census output area level, so much smaller areas, but my guess is the vast majority of the data would have got suppressed at that level.

I: Right, which would…

R: Because the geography is so much smaller, so in some ways, using a lesser way makes sense because of it matching the deprivation data, but we have to choose some sort of geography that was that kind of scale, so we could have ultimately used five kilometre grid scales, but we couldn't have used five hundred metre grid squares, so it's swings and roundabouts really, so we go to our lesser areas by default really and it also gives us the numbers. It's not just about suppression; it's about having my childhood obesity rates in a lesser way is twelve percent. I want that to be based on a reasonable number because if it's based on a very small number, then it's… even if I'm allowed to have the data, it's subject to a lot of random fluctuation.

I: Yeah.

R: So a lesser weight population is about fifteen hundred gives us a stable base for a reasonable denominator.

I: Yeah. Yeah. Sure. So it wouldn't be that important to then wish I could be as granular as postcode level.

R: Ideally, yes, I think that we... there would be different things you could do if you had data at that level and if you were able to do it in a protective fashion where we were able to link all the data at that level -

I: Yeah.

R: - and we might do a slightly different study design. We might actually… if we've got individual level data, knowing where their postcode is, then we can actually use individual level data instead of this aggregate ecological approach, so high resolution geography is always really a good thing, but you've just got to have sufficient access to it for it to be useful.

I: Yeah. Yeah. And I guess also it depends on what you really want to argue or find out because of the level of establishment the useful relationship, I guess you don't need to.

R: Yeah, it changes your research questions basically.

I: Yeah. Yeah. There was... reading the report, one bit that I was not understanding because of my lack of technical expertise, when you… at the point, point two point two, you describe how coastal proximity locations were related at one kilometre resolution to compensate for remote locations. I wasn't… if you could explain to me that part [laughter] because it is interesting for me for what I was saying about understanding methods of comparing places.

R: Yeah. Let me just have a look at the report again.

I: Yeah. It's at the point two by two on...

R: Yeah.

I: Okay. I'm trying to find, but my computer is being very slow.

R: [redacted]

I: Yeah.

R: Okay, so it's basically just saying [redacted] just put a point, points all along the coastline at one kilometre separation and then just calculate for each super output area how close it is to the nearest point. Partly because the GIS doesn't like… I've come across this before in similar work myself. Ideally, we just use the coastline, which is just the line, and we calculate the nearest point on the coastline, but that's quite a hard thing to do in GIS, so we have to first of all turn the coastline into a sequence of points, and then, the GIS can figure out which is the nearest point, so it's just a work around really to get to that.

I: Yeah. But look, could you explain again? So what makes the coastline different from the mainland?

R: Oh okay, so this is [redacted] done work in England, so… that had suggested that people's proximity to their coast affects their health and wellbeing in different kinds of ways, [redacted] shown that people who live closer to the coast are more likely to get more physical activity. They're more likely to report better mental health. Other sorts of health and wellbeing outcomes, [redacted] done this using data for the English Child Measurement Programme, so using the same data and found that kids living closer to the coast in urban areas… I think it was urban areas. Not in all areas, but in some geographical areas, kids living closer to the coast were less likely to be overweight, so basically we were doing the same thing for Wales with… and the argument is that there's… perhaps the coastal environment, paths and beaches and so on, provides an environment that's conducive to physical activity and being outdoors more, so that might support a hypothesis that kids might be less likely to be overweight if they live near the coast.

I: Okay. Okay. Thanks. Yes. But back to what you were explaining before, so the reference locations, coastal proximity reference locations, are they… they are not the locations of the weather stations.

R: [redacted]

I: Okay, so now maybe I'm trying... maybe I'm getting what you're at. It's basically for every lower super output area, there was a calculation of how distant it is from the coastline -

R: Yeah.

I: - to account for that finding, that effect that you were just talking about of the positive effects.

R: Yeah.

I: Okay, so this is what the explanation is for that bit.

R: Yeah, exactly.

I: Right. Okay. Okay. Okay. In terms of situating this project with other projects you've been working on, so this situation where you have results that you can't really trust so much because there's been the suppression, there has been… you can see it's too drastic for what it was looking at.

R: Yeah.

I: Has it happened many times before? Was this particularly dramatic of this situation, this project? How does this compare?

R: Yeah, I've not had it be a problem to this extent before. I don't know why that is. I mean, I think data are sometimes suppressed or whatever, but I guess… yeah, it's just never been to this extent before, so when I've worked before with, say, hospitals admissions data, we might have counts of less than five or something like that suppressed, so… and we might lose a certain

proportion of the data and we have had that with a study that we never actually published, but… we were doing some geographical analyses. I'm trying to remember what conditions there was. There was one… we basically had hospital admissions for three different conditions and because of the relative rarity of the different conditions, they had different levels of suppression, so the ones that were more common were fine, but I think the rare condition, probably like twenty, thirty percent of the data had been suppressed. It was less problematic because it was… or it appeared to be much more random, but it still clearly affected our ability to draw conclusions from that study, so it has happened in other datasets as well, but this is by far the most extreme one.

I: Right. Okay. Yeah.

R: But having said that, it is just because of the way that we've gone about it. It would be entirely fixable if we linked the two data resources in the opposite direction.

I: In the opposite direction.

R: Yeah.

I: Yeah. Absolutely.

R: And the presumption... I mean, I think… I guess the lesson from it really is that on the whole you rarely get problems with having data suppressed and so on with environmental data, whereas you often do with health data.

I: Yeah. Yeah. One thing that I haven't asked you is how did you make the judgement that you couldn't trust it because it was too suppressed? Did you have a reference figure? Did you… what is the experience… well, how do you work it out that it's not…

R: I... I don't think so actually. I think it was more a… we had a bit of a discussion between myself and the other collaborators. I think it was just clear to us looking at how much data had been suppressed that we couldn't really make use of it. I don't know at what point we might have said it was okay. I'm just trying to find here… see if I can find… oh yeah, so we've got different levels of cut-off, so it does… it ranges from twenty percent of the data being suppressed up to fifty five percent of the data being suppressed, and I think because we knew… because of the nature of the data that all… pretty much all of those suppressed data points were going to be low obesity rates, we knew that would just scupper the analysis really so… well, we did the analyses, but… yeah, I'm trying to think if it actually…

I: Yeah.

R: Yeah, I think we just thought that there wasn't much point in trying to chase it.

I: Yeah. Yeah. You did... so you couldn't, in the context of this project, produce (unclear 00:24:18) analyses from the other end, from… within SAIL, for example of what it would be if we... you couldn't do that, is that right, because you didn't have time or funding for it.

R: That's right.

I: Yeah. So you don't have a proportion of what's the effect of this suppression?

R: No, I don't think we could tell really. Yeah, and I… it's a good question because I don't know… we didn't really deal with this issue… we didn't recognise this issue until we had the data all sorted out. We didn't a priori set what would be an acceptable level of suppression.

I: Yeah. More trivial question, this project, so you collaborate with the others all remotely, did you need to meet up physically?

R: We had one meeting at the start of the project [redacted] but apart from that, we just collaborated by email and phone calls.

I: Right. Yeah. Yeah. It's... I guess also it's because you had all your pieces... all of your set pieces and your experience to people, so you can do…

R: Yeah, [redacted] I knew from a couple of conferences and things, so we knew each other already.

I: Okay.

R: So I think that helped.

I: Yeah.

R: So one project meeting in person was enough.

I: Okay. Okay. Yeah. Great. I think this was all I planned to discuss.

R: Okay.

I: It was more understanding the kinds of reasoning for, obviously, an inexperienced person like me, I'm a little bit behind all of this, the interpolations and the issue of the suppression stands out obviously in the report as being particularly salient and salient also for the kind of stuff that I've been interested in recently is to try to understand how secure systems like SAIL or MEDMI in what ways they can enable the reuse of the data at a sustainable and repeatable level and so on. Yeah. Thank you very much for your time.

R: Right.

I: Thanks very much.

R: Yeah, no problem.

I: Okay. Cool. And so if you remember to send me the thing and you remember there's also that opt-in option to make us able to release the transcript as a… as part of a collection on its own as open data and of always anonymised as regards with the others, and then, there's the second opt-in thing, whether you want to be identified, but it doesn't

mean that the others we mentioned would be identified. Obviously, in a small project like this, anonymity would be particularly difficult to ensure, so… but….

R: That's fine. Nothing particularly controversial in there. That's all fine by me.

I: Yeah, I don't think so. Cool. Thanks very much [redacted]

R: Yeah, no doubt [laughter]

I: Okay. Thanks very much.

R: [redacted]

I: [redacted]

(End of recording)