Tokyo Workshop on Genomic Medi...eutics, and Health 20240411-03

Thu, Apr 11, 2024 11:15AM • 33:38

**SUMMARY KEYWORDS**

data, icd, participants, records, codes, hospital, uk biobank, health outcomes, diagnosis, fields, nhs, date, registries, experts, disease, admission, verified, received, report, wales

00:12

So, as you see, in our presentations, one of the key strengths of YouTube is to be able to enter all that data we have the personnel data, the lifestyle data, all of the sample measurements and everything to ongoing health outcome follow up through the records. And as the population of our participants gets older, we see we're seeing more and more instant cases of disease occurring. And here you can see some predictions of the number of cases over the next 10 or so years, with particularly large predictive increases in diseases of old age. So, the data we have is very rich and has incredible values but has really to be set and it will be richer, but that data requires analyzing that data requires an understanding of the sources of data and considerations of how that affects how you should use it for your research. So we have five main sources of our data UK Biobank, four of those coming from linkage to national records, so hospital inpatient admissions, death registries, cancer registries, and the primary care data. So that's dental practice, doctor, and then we also have health outcomes data. Participants have self reported to us, either at the assessment centers for some patients. So the hospital inpatient data is known as the hospital episode statistics. And that's provided to the biobank from the NHS, which is the UK National Health Service and that's a publicly funded health care system within the UK devolved into four different regions, Ireland, England, Scotland and Wales UK Biobank in Northern Ireland. So we're receiving data from three different systems in England, Scotland and Wales. On the NHS is provided without charge people who are ordinarily invested in so that's all primary and secondary, except for dental and optical payments. All people use NHS have an NHS number, and that is what we use for linkage to our participants. So participants, I provided us with NHS number and that is linked to to the data, and that means that we have presumed in all of our participants, we will see that around 450,000 of our participants have a record or hospital episode in this says around 50,000 participants and that's most likely because they just haven't been tested. So we're going to tie in the price of hospital inpatient admission, in the time private. The vast majority of the care of care in the UK is provided by NHS. Services provided by so this was the details provided by different bodies for England, Scotland and Wales and England as provided by NHS England and Wales Clydesdale database managed by NHS Wales and in Scotland by Scotland, to Scotland. And this means there are slight differences between the data received from this provider. So that includes the dates of availability. So data is available earlier in Scotland and everywhere else starting in 1981. English data is available for 1997 Whilst data is available for 1999. So you can immediately see that our considerations on the hospital data is that the media Bureau slash profit, isn't it. So this is not going to be a good source of data about pediatric diagnostics. What types of data other statistics receive information or diagnosis that a participant has received in supplementation admission? So we'll see one primary diagnostic along with a variable number of seconds. So, the main diagnostic variable indicate the reason that a stepfather been admitted to hospital and a secondary data source appears in the other or underlying conditions that admission so you might get a participant who has had a stroke, but also has type two diabetes for example, and see the stroke and the main diagnostic criteria and then type two diabetes as well as the secondary type. And these are coded according to the world the World Health Organization's International Classification of Diseases, the majority according to the ICD 10, but the older sources data or use the ICD nine. There's also information around operations or procedures that patients have participants have undergone while they're in the hospital coding according to the MCS, again, all the sources data is recorded in s3 but the majority of the data is useless for me it also contains both main and secondary operational procedures for England and Wales only. We have information on maternity. So So, antenatal and birth information are for participants who have given birth and also episodes related to psychiatry. So information around the legal status of the admission for example. What data does the hospital episode statistics not contain? It doesn't contain information about the medications that someone will receive it for a few medications there's very high level information included in the operational procedures. So for example, there is a code X 35 Three intravenous immunotherapy but it doesn't tell you about the actual drug that was received. On this data, we have all types of the codes provided a valid that they are included within the ICD or within SAS, for example, for the kind of, we have a light touch policy on this so the researchers can really delve into the data and consider that for themselves. So how does the coding work? This example of the ICD 10 coding it's a hierarchical classification system containing 80,000 codes across 22 chapters. The chapters contain codes for related conditions and subtypes and other progressively sub classified. So for example, here in chapter one, you can see that there's cause infectious diseases, the first nine foods for the intestinal infectious disease, and then it's progressively sub classified until you're going into more and more specific intestine, intestinal infections. And this is very similar for the PCs. So again, progressive sophistication of operations procedures. in the UK by by this is data is available in two ways. So the first is as relational data tables, as soon as a master data table has been, which provides the information on the on the inefficient episodes pass or misuse of details like the admission discharge and the type of episode and each of these will have the participant identifier, the ID and then instance index. So here that would be the kind of numbered episode episodes that we've had on record for the first one, zero and one, two, etc. And then that will be linked to all the other tables so dive for diagnostic codes operations and procedures has been critical, which contains information about the sewage treatment, so like a number of days of cardiac or respiratory support that was received through a psychiatry table containing the administrative aspects of the service related to psychiatry, the procedures or diagnosis will be contained within those tables, but administrative in this table, and then the maternity and delivery tables containing maternity and birth. We also have some fields that have been made, summarize, made available as part of the main UK Biobank tabular data set. And these are particular for the diagnostic and operations. Surveys provide all the room codes for a particular participant and the first date that that occurs within the hospital records. So these would be sufficient if you'd wanted to know for example, just whether the participant has ever had that food and maybe when they were first diagnosed in hospital, but not the number of times or not the any subsequent repeat occurrences of this. Also still be careful with dates, as we say, as I said this as I would have liked actually to cleaning on this. So if that date is really important for your research, it might be a good idea to go into the records to have a look for yourself.

09:53

So what considerations do you have what things do you need to consider for the hospital data? And to consider that there's been some changes in Scotland of the versions of the clinical classifications used? These consider that we get this data from the NHS and that they may change the details are provided. Over time. There are changes to the ranges of values or preschool codes over time. So although personally, it uses ICD 10, that does go through additional changes and find the revisions. So for example, recently I was dealing with a query with a researcher was wondering why they couldn't find the code for fibromyalgia is 2012 and I had a look and solution systems and found that actually, that code was added in the fourth edition of the ICD 10 in 2012. So it's not surprising that it's not appearing in any of our participants prior when it was added to the coding system. Reporting Standards might change within data providers and guidance might be issued, so things may change over time. And also there are some slight differences in coding methods between the nations, particularly with the number of Secretary diagnostic codes that will be entered is different between the different nations. Each region has its own coding guidelines and responding to new standards. And, in addition, in Scotland, the time episodes is different from England and Wales. This might not be a major factor in research, but these things are really important to consider, especially when you're thinking about sending diagnostic codes or dates. In really want to address records receive data on the cost of death. Through the history national death registries receive for NHS England and Wales and from the NHS central register in Scotland. All participants were flagged by just following the data type recruitment. So we presume that we have been fiscal participants as of September 23. We have over 44,000 participants and you can see here the five most common primary causes of death and I see these are also claims using the ICD 10 information on a date of death and derived age at death and their primary as the causes of death as well. And that data is available both as amended, and also as a record table. We also have information that we've received from Nash national cancer registries. So national press registries receive information themselves from a variety of sources so they'll see possibly data, but data also from hospices, primary care and Franssen treatment centers too. And those registries are again different in Scotland and Wales and it also changed in their reporting over time as well. Because registries amalgamate from different sources, it's very important to take care of where participants have multiple records. Again for the light touch here to do this, or 4g, that's it then when there's multiple diagnoses, this might actually be suited on between different providers. So for example, a slight change in the ICD 10 code so for example, for unspecified, breast cancer to ductal breast cancer, for example, or maybe a minor variation in the date, someone diagnosed two days later at another airport. That doesn't mean there's going to be diagnosis that just means there's a kind of duplication of different providers. So although we would say the cancer registries here are probably the gold standard for dealing with them whether someone has been diagnosed with a cancer, we'd recommend especially to see with the number of distinct diagnosis participants received. This data is available as fields in the main data set and we received data details and circles from both before and after recruitment and so first, I'm transmittable the transaction from 1957 which is when registries were first established, of course earlier consistent with the fact that we the types of data that are available are the type of cancer coded with the ICD ICD 10. B n, the date of diagnosis and the age of cancer diagnosis and then also the histology and behavior of the tumor, which is coated with another International Classification of Diseases

14:21

describing that inflammation.

14:25

What data do we have available here? We don't have information on the status of the tumor. We don't have a tumor mutation details and details about what's happened in the therapy process. We know that that data would be really valuable and we are trying to make every effort to link to that enhanced past data. We don't have time for that the moment that we know the work

14:54

so also, we asked participants about their health at the assessment center. So I used to talked about this a bit earlier. But participants were asked to come to the assessment center with a list of medical conditions and approaches they've had in the past to help with memory and completeness. And then they were asked to report the instrument. snare and the knees were verified during a Bible interview. They were verified by a doctor. They don't provide a fully comprehensive health history, but they do reflect the vast recollections of the participant. So they provide us information on their illnesses including cancer operations and procedures and also any client medications that they will take. They will ask to provide either an age or year of the diagnosis operation and then the corresponding Angel yesterday interpolated for the field but didn't provide that information. We also do actually provide the information that the participants entered on the touchscreen questionnaire but we recommend that you use the verbal interview data because that is the one that's been verified and discussed by the treatment assessor. So why that might make that information be useful. It would be useful for information on pediatric illnesses for example, are on fire to win a raffle ticket and it would be useful for information on elicits less likely to be captured in especially secondary practice but all records and also some of the questions are answered in separate sections of the touchscreen questionnaire are useful for health outcomes that are maybe less likely to be factored in. So things like mental health and pain and then also things that are not covered by the NHS like dental and eyes. In addition, some of our follow up customers include questions that might be relevant to health outcomes. So for example, the work environment contains information on some respiratory or diagnoses and some risks and medications. The mental wellbeing questionnaire contains information on the self reported medication for mental health conditions. And the experience of pain includes questions about various medical conditions that cause or exacerbate pain. And again this kind of ease are two to add to some information that isn't factored within hospital records. And of course, the news data is only available for the participants who answered the questionnaire. We invite all participants who only have an email address for and that's around 3000 participants and returned to get a good response rate around 50% or varies based on the questioner. But as you said, local shootings or any of these I find the number of participants and then finally, we have the primary care data. So in the UK health setting, the first point of contact for health concerns is normally in general practice, primary care for non emergency situations, and some illnesses might be managed entirely within a primary care setting. So for diseases that are first diagnosed in primary care so things like dementia or type two diabetes includes the data allows for identification of more cases and earlier. So in this picture here, they showed that inclusion of the GP data approximately double the number of dementia pieces that are fractured according to the Dental Data sensor Indeed, many of these were then made to factor in statistics for this deep data and identification of a median of two and a quarter years earlier. However, as we talked about, this talk on Monday, there is no national system for collecting or sharing data. And so that means that the data currently has been able to obtain this for about 45% of the cohort up to 2070. And again, we really recognize the value of this data we know that it's really important and so there's lots of ongoing work environment to tie it to the types of data included our clinical events, prescriptions and registration records, and between different primary care practices that have different competing spies. Usually, they contain different food schema for both clinical and prescription data. And it's purely schema can be relatively complicated. We have a lot of resources available to

19:42

you to be able to try and break down into the schema without

19:52

again, I must have read my thoughts you need to make personal considerations when you use this data, but especially for the data. This data is real world data that was collected for administrating reasons it wasn't active for research and as I said, the classifications and the details can vary between different suppliers and different primary care practices. It's still got the potential to be extremely useful in health research. But you really do need to understand it well and you do need to consider the completeness of the coverage. The differences that you might see time and location to different local procedures that are very long. The procedures for curly hair are definitely not as standardized as the hospital reporting, and they're quite big differences. There can be differences between suppliers, and we'll just move between the general practices and house. Generally if you have SQL moving to a different area of dental practice, and they might mean that their records and tax advice different suppliers and food system or it might even mean that they've moved from a area that we have data for to an area that we don't see, so I'll just talk about some of the kind of healthy fields that you give our clients and provided. So we have these first occurrences field that have been created by mapping firms and cell support data to ICD 10 codes and combining the primary care hospital deaths and self report data together. So these map codes from other classification systems to three character ICD 10 codes for all foods except for external causes and cancers. Meaning that there's there's information available for 1200 ICD 10 sites. And then that means that we make available a field for the date that this was first reported in any source and then the source that that first record came from. So for example, for a CVP 25, which has gastric ulcers, you'd see that many different primary tests are repeated CT free post Mac to This PC, there's a self report mounting and then also, there's the ICD nine mapping to the ICD 10 as well. And then that will be made available in a safe place field where you can see the different sources of data that's available from and that can be linked to a related field that provides the date. So as always, there's considerations to be made to the intended as a first pass for health outcomes. No firmly Packard's can be unambiguously maps to the ICD 10. So the the mappings are incomplete and that's the same for the ICD nine even for all codings those that partly map to the ICD 10 Or that more than once repaired ICD 10 codes are excluded. And mandated protocols have been used as they've been downloaded from source and from the NHS. And these are likely to be debated by clinicians and experts and should be considered and verified before they used so again, another example, which has come from a receptivity. There are 5208 participants who have self report only as the source of their myocardial infarction report, which seems surprising. You'd expect that to come up in their hospital records that consider that the self report answer that's being prosecuted type 21 is heart attack slash myocardial infarction. And that could possibly have a few other meanings. Participants may have understood to mean different things, even though it's been verified in that public interview. Maybe in hospital, this is coded as identity. So my particular function, which doesn't have any more friends from the self report data, or maybe as I've been involved on the ischemic heart disease as well, also does not have any mapping to myself or data. That's just my hypothesis. You might be able to come up with a better one and I've not tested it. But you know, when you kind of come to these anomalies or unexplained things in the data, you will be able to go into the self report data card into the record level historical data. See as someone who has self reported a myocardial infarction actually has it, coders meaning something similar, but that's not it anyway. And that's just an example of being careful with these fields and considering your data sources rather than just kind of funding in it's also have some algorithmically defined outcomes which are very similar except that there are certain parents for those health conditions have been kind of selected and verified by some receptors. And the information about who was in those committees is all available in the resource so I'm showcase this uses the hospital access statistics based on support and death. It doesn't use any for primary care, again, provides fields that contain the date and the source of the basketball. So this role in these health outcomes here. We also provide course lists when you see what the features so the final thing about health outcomes data is that we also have the most expensive version of the UK Biobank research has been translated to a more common data model, which supports about advertorial Monday as a follow up how can they be standardize our data with other datasets is provided as a series of relational data tables. And then we also have some data which is specifically for COVID-19. Vaccination data, as well as the results of civility and studies which are

26:00

carried out during the pandemic.

26:05

So, again, what's what do you need to think about when you're thinking totally about your health? What how can you define your house? What sorts of data are most appropriate for your house outcome of interest? Some clinicians might be unlikely to require hospital inpatient treatment, they might be especially unlikely to appear as the primary diagnosis for hospital inpatient admission, but maybe that in the secondary underlying conditions How crucial is that? Date of diagnosis information? If it's a is very important, you might not want to use the summary fields because they use the wrong thing data. You might want to go in and have a look at it and see if you think those dates are recorded. is the first day of diagnosis. is in foster care, but there's no surprises. No summary. Hospital fields are talking about the first date of diagnosis in hospital if your disease is likely to be diagnosed impairment. That's unlikely to actually be where the disease was first recognized. Does your research require all information, information on all of the occasions when the diagnosis was recorded or just the first? If it requires information on all of the phases, you'll need to access the record level data? Are you interested in the primary diagnosis and the reasons are you also interested in interested in a secondary diagnosis? What coding list is most appropriate Milhouse? Up from the freeze of interests in a funny look in the literature consult with experts to think about the different tunings that might have been used across classification systems for your health outcome. And also, location considerations. As we've talked about, the National Health Service maybe isn't quite so national. There are differences between England, Scotland and Wales, so think about those as well. So, my tips for doing it, browse the health outcomes fields on showcase mean, don't just search the ICD 10 code or the name of the disease you're interested in in the search bar, you're quite likely to end up confused, particularly if it's a very broad thing like diabetes, read the resources available, a suit that are associated with the fields and categories you're interested in. Look at the documentation. Try and really understand the data. Think about that. Let's talk with experts. They can help you understand it. And then also, if you have a problem with the data to for example, the notes or resources, don't leave it clear and don't have the information required. Or you think there's an anomaly in the data that have explained. Please email us and we will make every effort to investigate them. And also, again, as I said, In my previous four

28:57

if possible.

29:15

So we have been working on the data for a while. So the first question is, it took us some time to really parse that data. So I was wondering, like, yeah, we got some tabular patient via Wi Fi, you guys just provide that data. So that, you know, like our scientists think that table and the number of columns in ourselves, so

29:42

you mean like to so listen to my father health outcome? Well, I mean, you know, we provide an awful lot of data and we probably experts in every single disease that is listed in that in that ICD 10. So I think that some of the first occurrences and the algorithm, especially a kind of attempt to make a something to make an attempt for that to happen to provide a list of without ICD 10 codes, but we've just used the mappings that are provided by the NHS like clinicians experts in so if you're an expert in stroke, you might have a different opinion on what those are. The algorithmically defined outcomes are maybe more verified by experts. And I think we would like to expand that and continue but it's very difficult for one person or one or even one organization to be an expert in every single thing.

30:43

So we have been working on the other side of the best effort. And the last one is called striker the mimic provided by the Fisher net. So they also have about 50,000 subjects from ICU, and then they provide different modalities besides the VA laboratory test. So each modality has a very broad format and tabular so which are the data becomes the standard benchmark for people doing AI in house? So maybe it's just my suggestion or common would be nice to process that data in a way that not like people are in CS or just, you know, different mindset and kind of have a different way of approaching the problem. Sometimes. It is a part of the process that will just shy away from somebody else to help them because they don't have the expertise. I said a lot of you guys so

31:49

yeah, it's a really simple one. I think we should look at that data set sample that you have given the time of collaboration with experts. In those meetings, to collaboration between

32:08

lot of questions I can ask you maybe one more burning question. So we work on like yes, I sort of lost, I thought was carried as long as I understand it's hard to keep monitoring the same such that over time, as we will have to you know, say your regular right. So we basis all the hospital Mondays and you have to wait for that. But what about going back to the past of pain, the past record for that same subject?

32:38

We do obtain all the data that is available. So for the hospital records, that is what is available. So it goes back to the 1990s the console goes back to 1957. So we have obtained all the records that are available from the participants prior to their recruitment.

32:55

Because from what I know that this transactional so we don't have the timestamp are the same subject or maybe they do build the hospital within a year or several times and you have different set of

33:09

data. Admissions and hospital we can talk further later I can show you the article.

33:26

Question