Appendix 2: Health Data Research UK Community - Ideas on how we can put the UK at the forefront of data science

We know the meeting on 20 July 2020 convened by Health Data Research UK and Lord Prior will provide an opportunity to understand what has changed during the COVID pandemic and use this to set a new and even more ambitious direction for health data research in the UK. To help inform this important discussion we invited people across our wider community – patients, member of the public, a colleague in the NHS, academia, industry and charity – to submit their ideas and suggestions in advance via our website. Using a range of our communication channels, we received the following responses.

	In relation to health data science, what have we learnt from COVID-19?	Over the next five years, what opportunities could there be for UK health data science?	Where are we now?	What needs to happen for us to meet the opportunities you identified?
1.	The need to clearly give a lead to both the public as well as the government. This is a great challenge since each group as its own agenda.	The opportunity to create an atmosphere of trust in the use of data science.	There are areas where the trust and transparency is lacking due to political purposes.	There needs to be a greater understanding on how advancements are used especially in the light for urgent real time data.
2.	That people work together well and fast when all focused on a problem that is everybody's. That our data infrastructures were woefully inadequate at the start of the pandemic to enable rapid insights. We have the data, and much of it is collated nationally but the infrastructure to support wide access for high priority research wasn't in place.	Retain infrastructures created for COVID-19 purposes to serve the generation of rapid insights from data post COVID. Move from the critical blocker problem being one of data access to address the challenges of data quality, improvement, scalable analyses etc. Improve international connections and global data sharing infrastructures so that they can	Now starting to seriously develop and scale the infrastructures in the world of health data that we've needed for years but need to focus on getting them solidly in place across the whole of the UK and developing them across all relevant structured data sources, simplifying access and including unstructured data.	Promote collaborative culture among academic institutions (abolishing REF would help a lot, as would focusing on value outputs for money in rather than simply seeing netting in large amounts of money and spending it as successful in itself) and between sectors (NHS-academia, industry) Insisting on a rewarding team science approaches (for academics

	That government does not understand that scale (e.g. 100,000 tests a day) is not the same as usefulness (e.g. ability to connect those tests to health records for clinical purposes or other healthcare datasets for research and planning) That none of us need to travel nearly as much as we have formerly - remote meetings are fabulous for many purposes and a great leveller. That rapid public and patient engagement in emergency response settings is possible and should be encouraged rather than avoided.	be called upon for the next pandemic or other global healthcare crisis Extend the current COPI powers to cover other healthcare crises that we have sat on for far too long - obesity epidemic/pandemic, mental health crisis, the public health burdens of cardiovascular diseases and cancer etc....these also kill millions and merit similar approaches. Extend scalable infrastructures for data in UK to include unstructured as well as structure data (e.g. images, free text etc....a myriad of buried information, currently highly fragmented). Retain the highly collaborative ethos developed during the COVID 19 emergence as it reduces unhealthy inter-institutional competition and rivalry and enables far more efficient, transparent and collaborative ways of working.	Improving UK-wide collaboration but still a way to go (old habit die hard and easy to slip backwards if we are not highly proactive)	abolish first and last authorship model, list contributors alphabetically and note everyone's contributions to all outputs / impacts. Celebrate inter-disciplinarity
3.	We need to make more use of the data that is collected and act fast	More collaboration across trusts and primary care to deliver more cohesive, consistent and cost-efficient care that results in better quality of life	At the beginning, we have the tools and the vision, we can see how it could work, but we a have long way to go.	Less projects that overlap consuming time and money. More focus with clear objectives and ensure meaningful and beneficial results that really will contribute to the future of the NHS digital evolution.

4.	There is no doubt now that we can create very valuable insight and solutions to health crises using existing health data and we can do that at speed. However, the value is dependent on the quality and diversity of the data and current gaps in data have been laid bare by COVID-19. We have learned being the first to publish is not everything. Time taken to engage with peers and the public still results in better quality results. We have learned that it is entirely possible to engage with the public and still move at speed.	To review the quality and diversity of current health data and take steps to improve it now to close the gaps. To run studies that cross existing health and social care boundaries to look properly at all determinants of health and help plan for the future. To build a new relationship with the public about use of health data and really engage and involve them. To build trust and understand where research is needed to help meet the needs of the public as well as the system.	We have the beginnings of the insight needed to do things better. We have seen what has worked, what hasn't. We have seen how public involvement and engagement can adapt and move at speed. Health data science has a public profile now that it has never had before. We have seen new relationships forged between organisations, nationally and internationally. But all of this is still new and could easily be lost again.	We can't afford to let practices slip back to the way they were. The OneLondon citizen juries show clearly that Londoners are supportive of the use of health data in research but they want to know what the research is for; to understand the benefits; to know they come back to the NHS and the public. This needs to be taken on board across the sector so that researchers understand that they need the support of the public to do that they do. If this new relationship is created there is huge potential for health data research to make a difference to the lives of people. This new relationship has to go hand in hand with investment in closing gaps in coverage and diversity of data. Without that immediate investment, research will only increase inequalities.

5.	We learnt that actual AI needs historical data, and when it doesn't have it, it fails to deliver and to have meaningful impact. The historical data is a human classification of the reality, transforming the fundamental progresses on AI into supervised learning (deep learning just as an example) where for each interaction done the goal is to remove as much uncertainty as possible from the data in order to be able to use AI. However, we know that human learning it's fundamentally unsupervised and capable to manage uncertainty. Probably this is a limitation of the AI binary world, however my conclusion is that we need to do a lot more research on unsupervised learning and even on self-supervised learning where a machine can learn from the context self-data. Only a UK wide research like HDR UK is capable approach it and create the conditions to deploy an environment where it would be possible to experiment some of the	Over the next five years we will probably see different challenges on the health sector but two of them can be easily predicted. Firstly, the pandemic is here to stay and other similar pandemics will probably happen, and their effect will need a clear knowledge about how to manage it, being it from the biological challenges, available resources or planning strategies (the outbreak in China had an early detection by a machine learning approach not because it recognized a pandemic but because some patterns had changed). It was also possible to understand very early that flights were an easy way to "airborne" the virus and spread it. This management has, from my perspective, two fundamental components, the capability to learn from the data and to produce knowledge from it. No human team (visible on the actual problems and challenges on contact tracing) can be compared to the computational power to deal with the data whoever I'm sure that actual AI don't have the tools on how to produce knowledge from it (visible on the Apple/Google Bluetooth where a lot of false positives were happening because of the missing awareness about the context-you can be next to someone but in another apartment separated by a wall). The knowledge will result from different layers of	I think that this is the easiest question as I think HDR UK has built the foundation for data sharing and integration which is a fundamental advantage.	I think that fundamentally two things can help to achieve the opportunities I've identified. Firstly, the evolution, of the substantial sites to environments centred in a type of research and expertise with the aim to explore further the challenges and opportunities and to create an environment for sharing science with focus on fundamental science. Secondly, the creation of virtual HUB's to address health specific problems bringing together existing structures in the HDR UK substantial sites.

theories. The good thing is that we already have the substancial sites to start with.

[1] Y. Hu, J. Jacob, G. J. M. Parker, D. J. Hawkes, J. R. Hurst, and D. Stoyanov, "The challenges of deploying artificial intelligence models in a rapidly evolving pandemic," Nature Machine Intelligence, vol. 54, pp. 1–3, May 2020.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015.

information whoever they need to be integrated to create an inference and reasoning environment.

Secondly, we are in a road to a more participatory health were the person (personalized medicine) will have tools (possibly apps and wearables) that will collect huge amounts of data, as self-reported or measured, that needs to be part of the decision making process on how better health can be managed and deployed. This will challenge the development of an architecture based on the AI capability to have self-supervised learning from data abstractions and to produce personalised profiling.

[3] M. Swan, "The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery," Big Data, vol. 1, no. 2, pp. 85–99, Jun. 2013.

[4] M. Swan, "Health 2050: The Realization of Personalized Medicine through Crowdsourcing, the Quantified Self, and the Participatory Biocitizen," JPM, vol. 2, no. 3, pp. 93–118, Sep. 2012.

[5] T. Fawcett, "Mining the Quantified Self: Personal Knowledge Discovery as a Challenge for Data Science," Big Data, vol. 3, no. 4, pp. 249–266, Dec. 2015.

6.		Uploading and integrating people's self-tracking information, of which there are probably millions (from their smart-phone step counts to fitbits to blood pressure readings, etc)	As far as I know (and I have been self-tracking for 40+ years, plus being part of Quantified Self) there is at present nothing done with this data.	A centralised system, similar to the national Cancer Registries, to invite contributions - perhaps in the form of .csv files, automatic uploads from their smart-phones - from patients
7.	We've learned that the supply of data for data scientists to make discoveries with and help better planning is inadequate, for the following reasons: 1. Interoperability: Even at a Trust level, just knowing how many beds had patients in at anyone time was a huge challenge. So then at a sector and national level, how can data science be used to innovate new solutions, aid planning, and monitor outcomes. Without a platform of accurate and comparable data, the benefits of data science were severely reduced. 2. Timeliness: Getting an up-to-date supply of information for data science was a challenge and highlighted how old data is. SUS is generally two months in arrears,	1. Direct patient care: Personal risk algorithms, remote monitoring, personalised care, drug discoveries, evidence creation and debunking at scale. 2. Population health management: Improved stratification algorithms moving away from antiquated and poorly evidenced systems and tools from the past 3. Secondary health uses: Create maps of patient flow, automatically changing pathways to improve wait times, divert to services with better outcomes, enable value based commissioning 4. Secondary uses: Research. Healthcare evaluation and research at pace and scale	Confused, bewildered, tired, yet optimistic!	Health data science in the UK is a user of predominantly NHS-created data. We need to improve the supply of this data in terms of scale, quality, and timelines to enable the real benefits of health data science. But this needs to be done with the public's consent, and I'd like to bring attention to the OneLondon consultation recently completed which debunked a lot of previous beliefs around healthcare data, and evidenced how with the right information given, the public are behind this. Further deliberations at different geographies is key to giving permission to use data better.

HES three. So data scientists couldn't get cracking on large secondary care data sets properly until 2-3 months into the crisis. 3. Equity of access. Some national bodies did not share/would not share in a timely manner (or at all) data they had collected and that there was an adequate legal basis to share. This resulted in data gaps or replication of effort to get data for data scientists to use. Data scientists outside of those national bodies therefore could not fully contribute, while I'm sure scientists within those bodies were over capacity. We've learned there is a long way to go to fully implement and embed the health data science into our care systems, yet at the same time there is a community of interest who are ready and willing to play a part and help to achieve many of the benefits and positive learning we did see.			

8.	How to form working collaborations very quickly when there is a common challenge.

Also, that we need to enable NHS hospitals to improve the quality of their data through enhanced digital maturity. | Internationally leading - our NHS provides one of the most complete health records for a population globally, and our ability to use this to enable real world evidence to inform and perform clinical studies should be world beating | We still have a fragmented health data ecosystem with vast heterogeneity in digital maturity. This is improving, and HDR-UK have been central to this, but much more should be done | We need to focus resource to help NHS care providers improve data quality and accessibility |
| 9. | The importance of researchers being able to access data that is linked, the power of research collaborations | Health Data Science to become the central component in relation to NHS, Technology, Commercial, Academic, Life Science moving forward for the benefit of our whole population | COVID-19 has brought health Data Science into the public gaze and is an important building block | To prove HDR is trustworthy in how data is used only possible through good public discussion, understanding and communication; honesty and transparency. defining of "fair value" is an important element |
| 10. | That it's possible to mount large studies quickly when some of the bureaucratic impediments are removed

That large well-conducted studies are more informative than lots of small ones

That citizens science (such as the Zoe app symptom study) can attract large numbers of participants and produce useful findings. | We can build on the Covid experience to conduct more large-scale, better studies of disease prevalence, treatment trials, health service use, and prevention. | We have quite a long way to go to create useable linked data sets that could be used to make a real impact. | Build public support for data use by ensuring that personal data is kept secure and used safely and that access procedures are clear and transparent

Encourage multi-disciplinary research collaborations that include patients and lay people

Work with NHS data custodians to develop common standards and access arrangements. |

11.	We reacted well to collating patient data considering the scale of data available. Our diagnostic capability is lacking. A more unified approach to handling medical data is required.	Crunch the data we have available to help predict and contain any future epidemic. Ensure better links between PHE, big pharma and our research community. Act globally.	We are still learning. I'm amazed at how thoroughly we have handled the data so far. This disease will have further impacts on our health in the future. Its quite apparent there are many more medical problems to come to light	A more integrated approach to research trying to remove the barriers to data share. Transparency is key
12.	The more data, the better the understanding and the quicker the treatment options. The acute need for organised data collection, storage, analysis and safety	Some entity must take the lead in convincing the public in the case FOR Data Science, establishing trust and security. With the growing use of AI algorithms, diagnostics will become more augmented and accurate. HDR_UK can be central in this role.	Still in the initial phase in most cases. UK BioBank has established a pathway in research terms for use of large datasets and multi agency collaboration. This needs to be expanded and developed.	A plan, timetable and finance plus a thorough public and patient involvement.
13.	That good collections of genomic, transcriptomic and clinical datasets coupled to innovative analytical methods provide the quickest insights into complex heterogeneous diseases.	Extend UK Biobank to be larger and more inclusive - it is a fabulous resource for world-leading science and a beacon for high-profile research projects.	We have initiatives that are well-funded (Genomic England and even HDR to some extent) that do not result in patient data being made available (with appropriate protection) to the wider research and especially SME community. This stifles UK innovation and the growth of a new generation of companies and researchers.	Make more data available for external researchers. UK Biobank is a fantastic example of what can be achieved.