UK Health Data Analytics Network
Research Roadmap Workshop Output

Healthcare Opportunities

In the first workshop attendees made their way around posters representing six Health Care Opportunity themes that had been identified in the Inaugural UKHDAN Workshop and the ATI workshop on health data analytics. Attendees were asked to comment on the themes refining the scope and importance of the theme and adding examples. The facilitator for each poster has provided a summary of the discussion.

1. Learning Health Systems

Data-driven, continuous improvement in health and healthcare for a given population, through refining and delivering best practice. Examples include real-time actionable analytics, rapid feedback, improved targeting, and safety monitoring.

Scope: Defining learning systems

- Relation between health systems and information systems is slippery/changing
- What is explanatory? Causes vs effects
- Training + changing cultures
- Needs consistent data resources
- Integration of data from different systems in real time
- Need appropriate technical support, underlying computing infrastructure for rapid, large-scale data analysis
- Capturing consistent data points
- Enabling data analytics
- AI vs clinical acumen
- Multiple 'levels of detail'
- The AA of medicine What do doctors do better?
- Concept of place
- Need for a proper model
- Standards: interoperability

Motivation:

- Joined-up
- Dealing with paper based records
- Learning best practice & creating consensus from 1000s of experts
- Creative & health
- An intervention that works in one population will not work in another population
- Improved decision making in practice eg risk score based in emergency care
- Avoiding false conclusions
- Quality assurance and transparency

Examples:

- Safety signals for pharmaco-vigilance
- Warning scores + performance feedback

- Precision/stratification medicine
- Changes to guidelines
- Public Health England – cancer targets
- Eliminating medical error
- Measure health outcome rather than activity
- Risk score based analysis in emergency care - sepsis breathomics
- Peer group comparators
- Providing actionable information at point of care
- Reducing time lag of research to practice
- Getting practitioners to own the metrics

2. New Insights from Integrating Non-Traditional Data

Gaining a more complete picture of individuals' health and patient experience by integrating and analysing 'big data' from diverse sources. Examples include health records, social media, mobile and wearable devices, patient experience sampling, digital footprints, and environmental data.

Repurposing devices

- There are many instances of devices being re-purposed to provide health-related data. It could be useful to produce a landscape of non-traditional sources and uses.
- Using a bike camera to detect falls and automatically call for help.
- The camera in a mobile phone can tell you about: the weather, ambient light, heart-rate (from changes in facial colouring), blood pressure from rise time on heart rate monitor.
- Retail loyalty card, gym membership provide useful data about diet, fitness etc.
- Interior positioning using Wi-Fi/Bluetooth from doctors' mobile phones is highly informative (e.g. about workload).
- Non-traditional uses of existing data (e.g. blood sugar, white cell count) to monitor health system performance
- Google search gives lots of information about individual
- Big Data Partnership looking at the tone of tweets to predict diabetes.
- Tone of voice on mobile phone. Non-verbal cues – for example, cough analysis for predicting onset of asthma.
- Patterns of mobile phone usage and mental state.
- Use of HMRC data to link loss of productivity to health problems (i.e. due to absence from work).
- Walking speed and health prediction.
- Ethnicity from supermarket purchases provides a context for interpreting data.
- Role of mobile phone masts in positioning, for example in Ebola crisis.

Data interpretation

- Environmental data (e.g. weather, toxicity, atmospheric particulates) provide important context for personal data. For example, remotely sensed images can give leaf coverage, which could be used in studies of wellbeing.
- Data visualization is important for presenting results to health care professionals and for model building.
- Aggregates of personal data can reveal new information that is not seen for the individual.
- Important to consider social influences as well as genetic in understanding disease (e.g. in studies of obesity).

- Lifespace diameter (radius of movements over a long period) informative about progression of onset of disease (e.g. dementia).
- Visualisation and exploration of data at different scales (special, temporal and people-count) is important.
- Really hard to link across datasets and then to maintain anonymity.
- Ginger.io app puts passive mobile data together (e.g. key strokes) to give behavioural profile (e.g. onset of depression). Raises issue of privacy and wider opportunities for detecting dementia etc..
- Notion of the Digital Phenotype in analysis.
- Important to link oral and general health data – currently silo'd
- Measure blood loss during surgery and relate to outcomes (e.g. recovery time).
- Extracting data from tissue removed during surgery, also from hair.
- Emergence of apps to solicit data through conventional methods.
- HSBC with Ernst and Young looking at changes in financial behaviour as a predictor (need to check the details on this).
- Importance of key contextual factors like ethnicity, mental state, and age.
- Opportunity to use patient experience data that is routinely collected by, for example, the National Rheumatoid Arthritis Society (check this is a correct example).
- Vast quantity of genomic data makes it hard to link to outcomes.

Human factors

- Can learn a lot about public views on healthcare from tweets – e.g. complaints about treatment in hospital.
- Role of advertisers in changing behaviour

General issues

- Providence and accuracy of data – for example, knowing whether the wearer has faked Fitbit data.
- Analysis needs to operate at different granularities of data: spatial, temporal, number of people.
- Regional perspective versus individual perspective.
- Preventative health is as important as diagnosis and treatment.
- The outcomes of data analysis are for policy makers and individuals.
- Inverse correlation of house price with obesity. Use this by putting obesity interventions in areas of lower hose prices.

Other issues captured

- One picture of lifestyle from body metrics, social media and GPS.
- Remember the Google work on tweets and epidemiology.
- Data fusion and the 'Curse of Dimensionality'
- In studies of criminality, bike theft increases in fine weather – are there lessons here for health data?
- Confounding factors?
- Most people think they look younger than they are.

3. Better Care through Patient-Specific Prediction

Using patient-specific models to guide the choice of the right preventive or therapeutic intervention at the right time. Examples include risk stratification, early warning to forestall relapse, and dynamic management of conditions.

scope: what topics should the theme include?

- Risk stratification (I.e. subgroups rather than individuals): patient-specific prediction is the extreme 'n=1' case of this and as such may not be achievable (or useful) in some scenarios.
- Economic evaluation of 'how much stratification' is useful – this is an optimisation problem where n=1/patient specific is an extreme. We would like to optimise the amount of stratification – trading off the 'costs' – additional data requirements, analytical time and expertise, danger of overfitting, with the 'benefits' – more targeted treatments, improved risk communication.
- Importance of 'defining' a subgroup: these may represent patient 'types', disease subtypes, treatment response or even patient preference.
- The danger of misclassification, dealing with 'outliers'.
- Ethical/equity/legal dimension: patients will receive different treatments by virtue of their stratification – is this ethical? Right of patient to ask for automated decision to be over-ruled.
- Impact on communication with patients - may be good or bad (e.g. explaining to a patient that they are in a subgroup that does not benefit from a particular treatment).
- Human-computer interaction: how does a clinician or patient interact with communicated risk?
- Capturing the value of patient-generated data (e.g. wearables)

motivation: why is the area important?

- It may inform clinical trials or other interventions that can be more targeted, or only relevant to certain groups of patients.
- Helps to tailor treatments, interventions and other responses of the healthcare system (although there is a risk of overload – e.g. responding to signals in monitoring systems that previously did not exist).
- Inform when the observational data tells us enough and when experiments (trials) are needed (perhaps n of 1 trials).
- Allow for the capture of environmental and other contextual information.

examples: provide illustrative examples

- A low-hanging fruit is personalisation of warfarin dosing.
- Enables us to exploit natural variation in prescribing to compare treatments (i.e. replicating an RCT).
- Early warning scores in paediatric intensive care

other issues captured

- Feature selection is an integral step in the methodology. On the other hand, we may also be able to identify which features that aren't currently known would be useful to
- know to reduce uncertainty (e.g. for this particular patient, test X would be useful to inform about Y).
- In a world of patient specific prediction and treatment, NICE (for example) would need to completely rethink how guidelines are written and implemented.

4. New Models of Technology-Enabled Care

Transforming care in the community by supporting patients and citizens to manage their own health, engaging carers, and reducing the workload of care professionals. Examples include data sharing to support self-care, co-produced care, and influencing health-related behaviours.

Features and applications

- Change of health paradigm
- Prognostic tool – contributing to research eg. Human phenome data development
- Generating evidence of benefit and defining outcomes – EVALUATION
- Reuse of commercial / ubiquitous data outside health realm eg. Clubcard for dementia; financial and debt data and mental health
- Important to use the technology as part of the system not in a BUBBLE
- Ie **CONNECTEDNESS**; if data is being generated and collected sit should be shared with other stakeholders and parts of the system
- Data visualisation relevant to the audience is key to making this useful
- Different models of care tailored to different 'types' of people eg well and unwell – would look very different eg cardiac rehab for unwell vs fitness monitoring for well.
- Enables the 'clinical care – informatics – dynamic team science' cyclic evaluation model
- Enable 'sensitive' monitoring – not overt and clunky – integrated into life (eg. clocks)
- Could help engagement – HALO effect
- Useful to ageing population and costs / care model
- Gives a fuller picture of a person move sus nearer to the wellness model rather than sickness model
- Gives patients ownership of care
- Allows coproduction of care with patients – key to engage people from start of lifecycle
- Nudges behaviour change
- Positive and negative unintended consequences could result – important to understand & monitor these
- Examples of application: medication adherence; smoking management; mental health; clinical decision support; personal health; brain training; Virtual Wards; smart homes; treatment scheduling eg. bloodcell monitoring at home
- Management of risk – stratification and prediction
- Routine data for health

Challenges

- The ‘creepiness’ factor of being monitored
- Culture and mindset shift
- Safety of design is key
- Individual differences- & power structures in health (paternalistic)
- Are we measuring the RIGHT THING?
- Could be a totalitarian model
- Important to engage all stakeholders – awareness of conflicting incentives across the system
- The ethics of what is done with data and feedback and how people may react to information given (eg. genetic counselling)
- Worried well effect – don’t reach the needed populations (hard to reach, at risk) but over worry the well – and could lead to widening the gap between social groups
- Access – denied to those who can’t afford tech?
- Change management is key
- Change to the health infrastructure is key
- Need to define who the data is for and for what purpose
- Emancipatory technology design vs. disempowering users?
- Psychosocialtechnical challenges
- Implementation of the system needs to be appropriate

5. Personalising Care

Transforming patients’ experience of healthcare, empowering them to engage in their own care and manage it in a way that suits their needs. Examples include personal virtual health assistant, individual goal setting, and personalised information and feedback.

> *Note: all information written on sticky notes is in green below, the narrative is all memory of the facilitator.*

At first we started with a discussion about what was meant by ‘personalising care’ with highlighted terms as follows:

- Co-producer
  - Less passive patients
  - Scalable
  - Prevention
  - Patient responsible for own health

The ideas being that healthcare and prevention could be scalable to the entire population if patients were co-producers in their own health and care. We focused on patients taking some responsibility for their health and care and being empowered by the health service to do so. This quickly led us to discuss prevention of health conditions being as important as looking after oneself with health problems.

Within this discussion the name quite quickly changed to “Co-Producing Health” – this was not challenged by others – although all were invited to comment on it. It seemed universally accepted. There was also discussion about how this was “Broader than care which implies medical intervention, what about prevention”?

More concepts on how co-producing health might work included:

- Access to own records
- Supporting personal agency
- Decision support for individuals [the patients or healthy public]
- Providing interpretation, filtering information for relevance
- Summarising information
- Crowd sourcing

Crowd sourcing was a popular topic, particularly with ‘People like me’ type of websites and offering community support. The idea was that people within the community could help one another.

There were ideas that this could lead to:

- Shared goal setting between healthcare professional and patients, common health concepts
- Enabling choice by patients
- Increase treatment compliance
- Addressing chronic low-level problems can the NHS afford it/ helping increase quality of life through IT and IT groups
- Community of people with similar situation- patients like me
- Reduce/ remove organisational barriers to personal health data
- Personalising understanding condition/ health
- Lots of useful information through questionnaire as opposed to genetic data

Goals included

- Enhancing wellbeing (Example: recording what you are eating)
- Empowerment, education, self-help- people like me, community based
- Mobile apps to monitor lifestyle
- Community agency
- Delivery of prevention at birth
- IT solutions to health literacy
- Medical solutions, social interface
- Surveillance by medical authorities of everything measureable
- Health literacy (multiple times)
- Individuals choose goals relevant to them
- Prevention activities
- Patients feeling in control
- Co-produced care

- Helping make It difficult to do the wrong thing
- Create incentives to do the right thing and barriers to do wrong thing
- Alternative care plans
- Accessing people in their own space without them having to access care
- Personal Agency, lived experience, personalised choice

Issues that we needed to address included:

- Redefining personalised
- What set of people are we addressing?
- Competing concepts of health and well being
- Views of healthy
- Representative data?
- Common language, translation of medical language, one size does not fit all
- Conveying information is not sufficient
- HEALTH LITERACY
- Limits of allowance of self-decision, who decides
- Mathematical modelling/ big data problem
- Unintended consequences of access to data

A solution offered to the big data problem and probability was PICTURES.

Concerns and risks expressed included:

- What does data miss, does it take away personal choice?
- Health inequalities
- Technology access inequality
- Unintended consequences
- Augmentation theory
- Technology is currently aimed at well
- Ethics
- Unintended consequences, misdiagnosis
- Risk assessment – Speigelhalter
- Health inequalities
- Could data be common currency?
- Could this widen the gap in health inequalities?
-

There was a discussion about whether or not this would lead to “Stopping the Goole phenomenon” or starting it.

At some point there was a debate about whether or not co-producing personalised care vs public health was more effective for the same outcome. This was not resolved and there were clear differences of opinion.

There was also discussion about needing “Clear terms and conditions in interface for access”.

6. Characterising the Human Phenome

Redefining disease classes to enable a better link between biology and medicine. Examples include phenotype/endotype discovery, and defining core morbidities of the elderly population.

Biomedical research and clinical practice build on disease definitions that are often highly imprecise, derived from superficial manifestations of disease, and poorly rooted in biology. Well-known examples are pulmonary disease (asthma is suspected to be an umbrella term for perhaps 10 or more different conditions) and mental health (schizophrenia and dementia are notorious examples). But it was also recently discovered that there are 5 different types of "type 2 diabetes mellitus" and 3 different types of pancreatic cancer.

These limitations in the existing nosology restrict our understanding of human pathology, hamper the opportunities for effective treatment, and reduce the efficiency of the treatment discovery pipeline. One wonders how much more efficient the Salford Lung Study would be if we had a better understanding of pulmonary disease classes. It therefore makes sense to redefine disease classes to enable a better link between biology and medicine, for instance by defining classes based on aetiology or treatment response. The latter approach is also key to effective precision medicine.

It was noted during the discussions that a purely biological approach would probably have its limitations, as there are also cultural and socio-economic factors that affect treatment effectiveness, e.g. through adherence. It was therefore suggested to identify for "behavioural phenotypes" as well. Furthermore, if redefining the nosology leads to more disease classes then there will also be more opportunities for mistakes and misclassification – perhaps nullifying the theoretical advantages.

It was also noted that disease classes have always evolved and will probably always continue to evolve. Recent examples are the redefinition of acute myocardial infarction to acute coronary syndrome, the recognition of cardiometabolic syndrome (cardiovascular disease plus diabetes) as an independent disease, and the identification of IgG4-related disease in rheumatology. Changes in our environment and evolution of pathogenic agents will drive the need for continued reconsideration of disease classes.

A more fundamental question is whether it is helpful to have disease classes at all. Some researchers have argued for a new paradigm that tries to describe disease purely in terms of observed pathology, allowing it to be different for each individual. While this makes sense from a precision medicine perspective, it is hard to conceive how the health system could still function without having disease classes. For instance, how would GPs communicate to their patients when they should visit the surgery, or what they have?

- Behavioural phenotypes
- Everything we see is biased
- Evolving disease classes
- More classes-more opportunities for mistakes
- No classes at all?
- Co-morbidity networks
- Key to precision medicine
- Efficiency at discovery pipeline

- Communicating difference in treatment between subtypes of illness to patients & GPs
- Treatment response
- Pancreatic cancer are actually 3 different diseases
- Mental health diagnosis
- Asthma
- T1-T5 diabetes
- Metabolic syndrome
- Schizophrenia
- Culture of socio-economic factors affect treatment effectiveness/compliance
- Biomedical research and clinical practice build disease definitions which are highly imprecise
- Identifying variations in outcomes from data – maybe need for subtypes
- What medically relevant conditions do not have a straight forward underlying biological explanation? Eg. Depression, ADHD, Autism.
- Acute colonary syndrome.
- 1664 disease (rheumatology)
- n – of – 1finals
- Microbiome effects on susceptibility and treatment outcome (eg. Respiratory gastrointestinal)

Data Science Challenges

In the first workshop attendees made their way around posters representing five Data Science Challenge themes that had been identified in the Inaugural UKHDAN Workshop and the ATI workshop on health data analytics. Attendees were asked to comment on the themes refining the scope and importance of the theme and adding examples. The facilitator for each poster has provided a summary of the discussion.

1. Integrating Heterogeneous Data Sources

New healthcare opportunities bring extreme challenges in combining disparate kinds of data of varying reliability, at volume. Challenges include data modelling, data provenance, robust linkage, normalisation, and managing uncertainty.

- Need to understand value of data
    - Where is comes from
    - Context dependent
    - Depiction question?

- Natural variability (e.g. expression) =/= Reliability (e.g. self report)
- Weighting by reliability
- Difference between integration & linkage of data
- Reliability of data mining – when to use it?
- Data Integration Systems
- Need to state this in terms of patient safety
- Sharing of scoping
- eLab

- Data journal
- Model for data collection processor
- Model ⇄ Data
- Qualitative & quantative data (disparate)
- Bias analysis
- Standardise coding?
- Context in which data is captured – part of data provenance?
- Uncertainty increased by multiple data sets.
- Visualisation
- Clinical data will always be flawed due to the nature of the data itself.
- Statistical inferencing
- Cannot define uncertainty without faster?
- Free text an explain inconsistency and help solve it
- How to collect standard health data so intergratable
- International Standards
- Experiments in data creation can help understand and address uncertainty e.g. EMISUS Vision
- Look at either disciplines using heterogeneous data eg. Via data shared entology/syntax
- Map ontology
- Look at other disciplines/ standards
- Multiple sensors- logic to find out what it means
- How mud data-driven technologies replace clinical data
- Reduce uncertainty by better interfaces
- Ned to understand data well - are differences due to scale of measurement or the population differences?
- Simulate virtual database/ capture missing?
- Data warehouse: integrate data of different sources
- Single person identifier
- Audit process
- Provenance chain or metadata
- Meta data important

2. Dealing with Missing, Unreliable and Corrupted Data

Issues of data quality are exacerbated by new forms of data, often collected under uncontrolled conditions. Challenges include robust inference from imperfect data, data imputation, and artefact detection.

Issues of data quality are exacerbated by new forms of data, often collected under uncontrolled conditions. Examples are routinely collected data in electronic health records (EHRs) which typically have many missing values; contain multiple, inconsistent recordings of the same item; and are based on measurement methods (e.g. biochemical assays) that change over time. Sometimes data is deliberately left out of research datasets due to information governance issues, such as clinical narratives. Furthermore, variations in care consumption (both between individuals and over time) limit the representativeness of these data for the general population.

It was suggested that some of these problems can be avoided by preventing data errors (e.g. recording of impossible dates) at the source. It could also be helpful to give feedback to clinicians about which data elements are missing that would influence the output of decision support tools such as predictive models. However, while both being useful approaches, they would not solve the problem entirely and would leave some aspects of it unaffected.

An obvious question that arises when considering these issues is: Which level of data quality would be acceptable? In general, workshop participants agreed that such a level does not exist, and all datasets have their limitations – even those collected using conventional designs such as controlled trials and cohort studies. There was broad consensus that we should accept the data as it is, and develop better methods to handle data of imperfect quality. Some of the solutions that were suggested were integration of different data sources, use of simulation methods and synthetic data, and representing missing values are intervals. It was also suggested to develop machine learning methods for data curation, by systematically comparing manually curated and non-curated datasets.

It was recognised throughout the discussion that a core barrier is the general lack of metadata, which leads situations "where you don't know what you don't know". Clearly more efforts are needed to design meta languages to express data quality and to develop procedures to record or derive such meta data automatically. This could also improve our understanding of the mechanisms that lead to missing, unreliable or corrupted data.

Generally speaking, data quality issues are not unique to health data, and there are probably things that can be learned from other fields (such as physics). However workshop participants recognised that some aspects of health data quality reflect strong human involvement that does not occur in other fields. EHRs are essentially a record of engagement of citizens with the health service: when citizens decide not to visit their doctor, their data will be missing. Similarly, clinicians decide which data are collected during clinical consultations. As a result, when they decide to not collect certain items, this is informative in itself. Finally, there is often an element of subjectivity in the data that is recorded (e.g. diagnoses). For these reasons it makes sense to study methods for dealing with health data quality separately from other data science fields

- Accept the data as it is
- Integrating data sources
- Big brother?
- Metadata about quality
- Prevent errors (e.g. wrong dates at the source)
- Subjectivity of collectors
- Individual affected by outcome
- Interest in life course data
- Deliberating missing data (e.g. free text EPRD)

- Social relationship (eg. Patient-doctors) distant measurements
- Human is the loop
- Missing data as interval
- Variation is care consumption
- You don't know what you don't know
- Stimulation methods/synthetic data
- Lack of meta data
- Better meta language for expressing uncertainty
- EBM myths
- Multiple inconsistent recordings of the same item
- Missing data in EHRS
- Wrong incentives
- What is good enough?
- Try to pitch more positively?
- Learn from other fields (physics)
- Changes in recoding methods
- Safety issues
- GP records are a record of engagement with the health service
- Keep the rubbish and use it to develop ML methods for data cleaning
- Feedback which elements are missing that would influence the prediction eg. When a clinician uses a CVD tool, say "recording this patient's ethnicity will improve the confidence
- More use of free text sources
- Increase understanding + mechanisms interval by x – do you want me to record this?"
- How to generalise from proxies
- Redefine nosology
- TV bias

3. Marriage of Human and Machine

New data-driven approaches to healthcare pose serious challenges in usability. Challenges include effective engagement of stakeholders (patients, carers, professionals) in design, data visualisation, filtering for relevance, adaptive interfaces, and just-in-time feedback.

> *Note: all information written on sticky notes is in green below, the narrative is all memory of the facilitator.*

There was a general dislike of the term "marriage" and the title was changed to "Human & Machine".

Throughout discussions in different groups we looked at what was central to human and machine. The following surfaced (in order):

- User Centred design
- Patient at Centre
- Training at centre

- Visualisation in centre! Partnership – this included using the system to create partnership between the patient and clinician
- While ‘who’ should be at the centre evolved, it was universally agreed that a “Data centred data driven system [was] wrong for the patient”.

Methods and implementation became a major part of the discussion, as these were viewed as problems requiring solution in the bringing together of human and machine. The following were raised:

- Hold the care model system level
- Fit with care model
- Better utilisation of the human component
- Requirement driven
- Emotional Labour coming together, hard stuff
- Don’t forget implementation
- Benefit to individual
- Build software that is user context sensitive
- Measuring mistakes? Adaptive system
- Slow change = slow acceptance
- Intuitive interfaces
- Interfaces between systems so that they can be flexible & different
- Patient input into medical record
- Learn from other industries
- NLP Automatic extraction
- Understanding machine limitations

We also discussed positive and negative attributes of machines and humans in this context:

- Positive human:
  - Empathy
  - Respond to unique intelligence and connect subtitles
  - Consent
- Negative human:
  - Ignore;
  - Judge;
  - Tired;
  - Distracted;
  - Errors;
  - Emotions -> decisions; gossip
- Positive Machine
  - Calculator remembers

  - No bias
  - Provides information for data sharing
  - Logistics supports patient memory consistently

- Negative Machine
  - May crash
  - No intuition
  - Not good at empathy caring or listening
  - Issue of security (hackers)

Barriers to this type of system included:

- Learned negative behaviour based on interface that doesn't work
- Barriers of Sharing across org. boundaries
- Poorly designed systems
- Pt acceptability
- Introduction empowerment communication mediating expectations
- Incentives
- Reliability of data, Machine mistrusting human

Solutions to these barriers included:

- Education
- Defining whose responsible
- Common Languages on user interface Doctor/ Pt looking at together
- Disagreement is ok
- Innovative training
- Getting the computer to be a barrier wrong or right
- Social interactions, Machines to support
- Learning to use tool training
- Machine compatible Manual system Building Confidence

This topic raised a lot of concerns related to patient safety, medical interpretation and legal concerns. The following were discussed:

- Machine should not provide decisions but information
- Lack of involvement of user
- System needs to explain itself
- Accountability machine vs human
- Accountability Trust vs mistrust
- Conflict who makes decisions & how?
- Unintended consequences: Letting system make decision
- Technology took over consultation

Ethics & Privacy were highlighted as critically important.

4. Characterising Complex Temporal Structure

Temporal patterns in often high-dimensional data have the potential to provide important new insights, but are challenging to exploit. Challenges include modelling complex often episodic behaviour, combining data sampled at different frequencies, and calibration drift.

General issues

- Temporal dimension brings with it
  - High resolution multivariate data resulting in very high dimensionality,
  - Often sparse and irregular sampling,
- Major challenge is to discover clinically important hypotheses from the vast array of available data, normally collected for an entirely different purpose. For example, without manually formulating a hypothesis about Netflix viewing figures and back pain, how would one discover a link automatically if this existing.
- There is a link between the 5 themes: need 1,2,3 to get 4, then can do 5.

Ensuring comparability of temporal data

- Changes in coding conventions over time within e-records.
- Lack of a shared ontology for episodic data (e.g. within GP records).
- Instrument calibration may be different between trajectories and may drift for a single trajectory.
- Clinicians collect different data (possibly driven by the e-record system used?).
- Need a good dialogue with those designing and managing data sources – don't rely on the raw data alone.
- Different sources of data may still be comparable within a unifying model. For example, is it possible to compare textual data from tweets, emails, facebook, and blogs through transforming (translating) into an underlying representation or inter-lingua, the problem being that words may by convention have different meanings in the different forms of social media?
- Learning from other domains where data is semantics is changing:
  - Models of the way in which the Arabic language has changed over the past 1000 years.
  - Aircraft designers need to worry about backward compatibility, e.g. with shared parts.

Statistical inference

- Dealing with critical events that are infrequent or absent within dataset.
- Dealing with relativistic measures. For example, can we infer that people are better off when all we have is that this group are in the top 10% etc.
- There is a difference between predictive and causal inference.
- The need to adapt/evolve models over time as the population changes.
- Need for frequency domain analysis to determine whether things happen at specific times of the year or day (e.g. cardiac events peak at 10am and 6pm).
- Loss of data from an individual may not be independent of health condition, hence potentially biasing models. Cohort-based datasets less prone to this.
- There may be many hidden variables.
- Abstraction is important in order to deal with high dimensionality – for example, mapping continuous data into a sub-space or onto a discrete variable (e.g. textual categories).
- When is correlation a valid causation – temporal ordering is an important constraint here.
- Adaptive sampling – knowing when to sample.

The importance of context

Contextual influences may not be explicitly recorded in dataset – for example:

- Rise in referrals for skin cancer at the end of the summer – concern over sun-exposure,
- Influence of storyline in Coronation Street,
- Kylie Minogue's' diagnosis of breast cancer.
- Circadian rhythms and the rise in heart attack rates around 10am and 6pm.
- Changes in the health system – e.g. doctors' contracts
- Social influences, for example through conversation with GP

Data visualization and summarization of temporal data important in exploring hypotheses that have clinical value – including for data that may have been collected for a different purpose.

Applications of temporal analysis

Early prediction and monitoring of cognitive decline.
Fundamental importance of temporal data in predictive analysis.
Often need complete picture over a long time span, requiring temporal data (e.g. for research on diet).
Importance within health economics
Importance of understanding the risk factors of a modelling framework.

Other points made, but not sure of relevance here

Importance of the temporal window
Impact of noise
Modelling episodic data – what is normal?
Lacking integration of methods – longitudinal data
Is it safe to abstract and quantize.
Models of the population versus models for individuals.
Importance of age differences.

5. Identifying Subgroups

Discovering structure within data is a key technology for precision healthcare, identifying subtypes of susceptibility, disease and response to treatment. Challenges include robust discovery in high-dimensions, dimensionality reduction, and feature selection.

scope: what topics should the theme include?

- Relative merits of unsupervised v semi-supervised v supervised approaches to classification/clustering.
- Transportability/stability and validity of identified subgroups to new data / contexts.
- Appropriate trade-off with regard to size of subgroups: smaller = more targeted care but larger = more robust.
- Membership of multiple subgroups simultaneously (multiple latent dimensions).

- Uncertainty in subgroup assignment (probabilistic) and the importance of this (i.e. how do we treat a patient when we don't really know which subgroup they belong to). The cost of misclassification may be high.
- Identifying homogeneous v heterogeneous clusters – i.e. some clusters may represent well-defined subgroups of patients while others are 'garbage collecting'

distinctiveness: why is health data particularly challenging?

- There is a potential clash between 'machine learnable' subtypes (i.e. those best supported by the data, clustering method used, etc) and those that are clinically meaningful. We may end up with 'new' categories of disease that may overlap/straddle existing ones.
- Distinction between prediction and causal inference. In other fields, prediction may be 'enough' but in healthcare we typically want to understand subgroups and relationships between variables in a causal sense. Methods for causal inference using observational data remain underdeveloped.
- Subgroups are a fuzzy concept and the 'right' subgroup may change depending on context. For example, a cancer clinician may be interested in different subgroup classifications than a dietician (in cancer patients).
- The challenge of creating multidisciplinary teams to build models to identify subgroups, and to interpret the output.
- Whenever 'automatic profiling' is carried out, the patient has a right to object.
- May also be SIMILARITIES with other areas – e.g. genre classification of webpages, jet engine monitoring, identifying themes in religious texts.

links: how are the five themes related?

- @4: temporal structure may be captured (or summarised more succinctly) in subgroups – i.e. subgroups can simply act as a label (latent variable) to summarise temporal complexity in a more manageable way.
- @1: integration of data from other sources (possibly at other levels e.g. geographic / contextual data) may inform subgroups.
- @2: missing/unreliable data may contribute to high uncertainty in subgroup assignment or misclassification.

Group Name: Health Data Pipeline	Colour: **Red**

Broader Issue #1: Data Catalogue

- What data are available?

Accessible to everyone

- Who is the data guardian?
- How to get access?
    - Restrictions
    - Prerequisites

Research Challenge #2: Methods, governance & ethics

- Access
- Linkage
- Data management
- Platform and infrastructure
- Data standards

Research Challenge #3: Engagement, dissemination & translation

- Dialogue with all stakeholders: patients, society, doctors, scientists, analysis, policy makers
- Ensuring that results of analysis are available
- Responsibility for translation

If you wish, draw a diagram which represents your group's area, below:	Group Members: **Lydia Drumright** Charles Taylor Arief Gusnanto Anna Palczewska Stephen Swift Kayleigh Mason Michelle Morris

Group Name:

Colour: Yellow

Broader Issue #1: Health Data inequality
(Trigg's Land?)

Inverse data law: the more health needs people have, the less likely they are represented by the data (both clinically and in research)

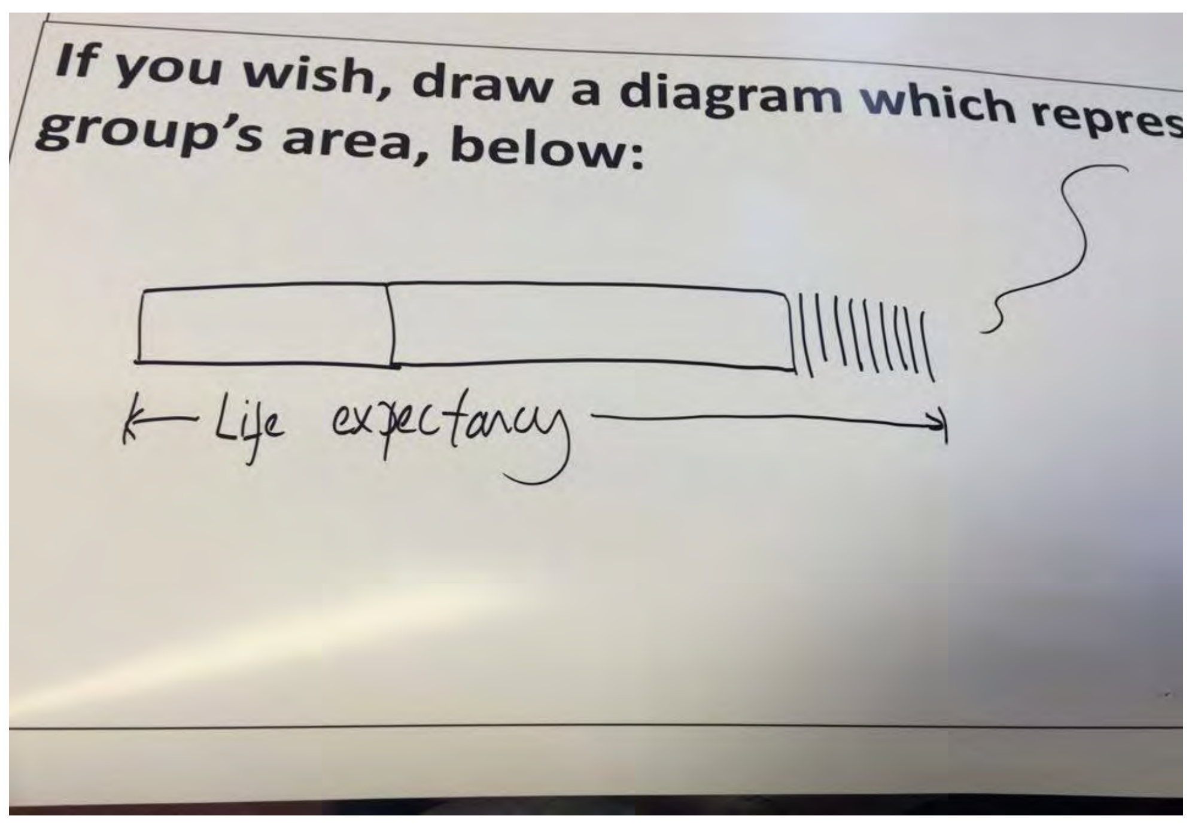
Broader Issue #2:Design and implementation

- Broad stakeholder engagement processes are ineffective self-selected cohort
- Implementation should exploit the adaptive nature of software

Broader Issue #3: Data literacy and transparency

Culture of exploiting illiteracy reduced trust in data science

If you wish, draw a diagram which represents your group's area, below:


If you wish, draw a diagram which repres
group's area, below:

Group Members:
Sarah Twigg
Niels Peek

Group Name: Computing infrastructure and enabling technologies

Colour: **Orange**

Broader Issue #1: Software

- Databases = domain specific
  - Non-relational data – eg images from proteomics
  - New database architectures to support this
- Data analytic software for Big Data Platforms:
  - Integration of algorithms with Hadoop, Spark, Met
  - (some are available, but lots not and lots very immature)
- Training

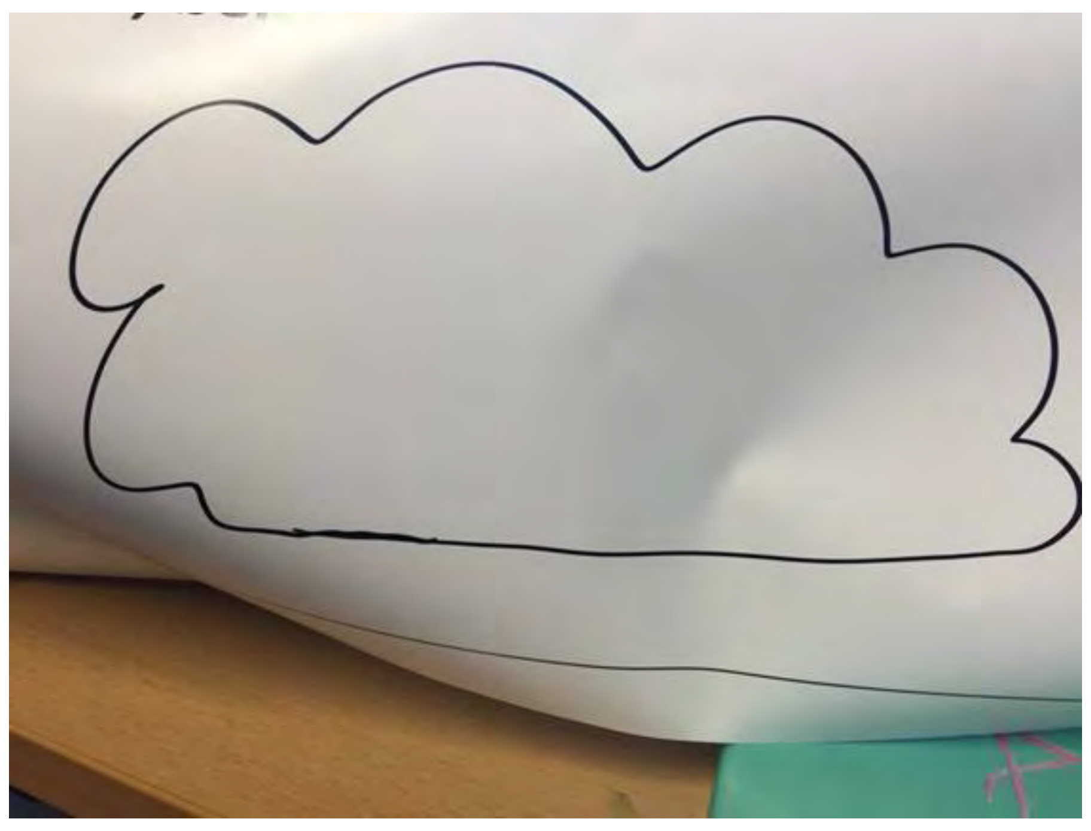
Broader Issue #2: Hardware

- Getting data to the right place
  - Interfacing with legacy systems
  - Non-standard data formats
  - Maintain security
- Continuity of resources when using external clouds:
  - Must be NHS compliant centre
  - No UK research cloud
  - What happens when project ends?
  - How to make use of ECZ or Agure
- Some need for local data processing:
  - ML (machine learning) on smartphone, on smartwatch feasible?
  - Sensor interface standards: shorter route from wearable to data centre
- Hardware infrastructure needed should be a part of the roadmap

Broader Issue #3: Skills and accessibility

- Re-using what's already available/generic components
  - Open source is necessary but not sufficient. Needs documentation, training
  - Interaction with MDD compliance and audit log of charges for CE marking who 'owns' open sources, sustainability? Model like Linux Kernel /Apache model?
- Balance between open source and closed. Recognising what's needed/works best in order to give a viable model for a business
- Training:
  - Career paths not losing skills, re-doing everything from scratch, embedding technical expertise (healthcare and research)
  - Short term ease of access getting up to speed with complex tools

If you wish, draw a diagram which represents your group's area, below:



Group Members:
Alex Casson
David Tian
Paris Yaipanis
Jonathan Tedds

Group Name:

Colour: Pink

Broader Issue #1: P.R for the field

Step change in joint understanding
Data science <-> health care and social data who knows both areas?
Describing the openness of ‘data science’ field
Career paths/horizons? future

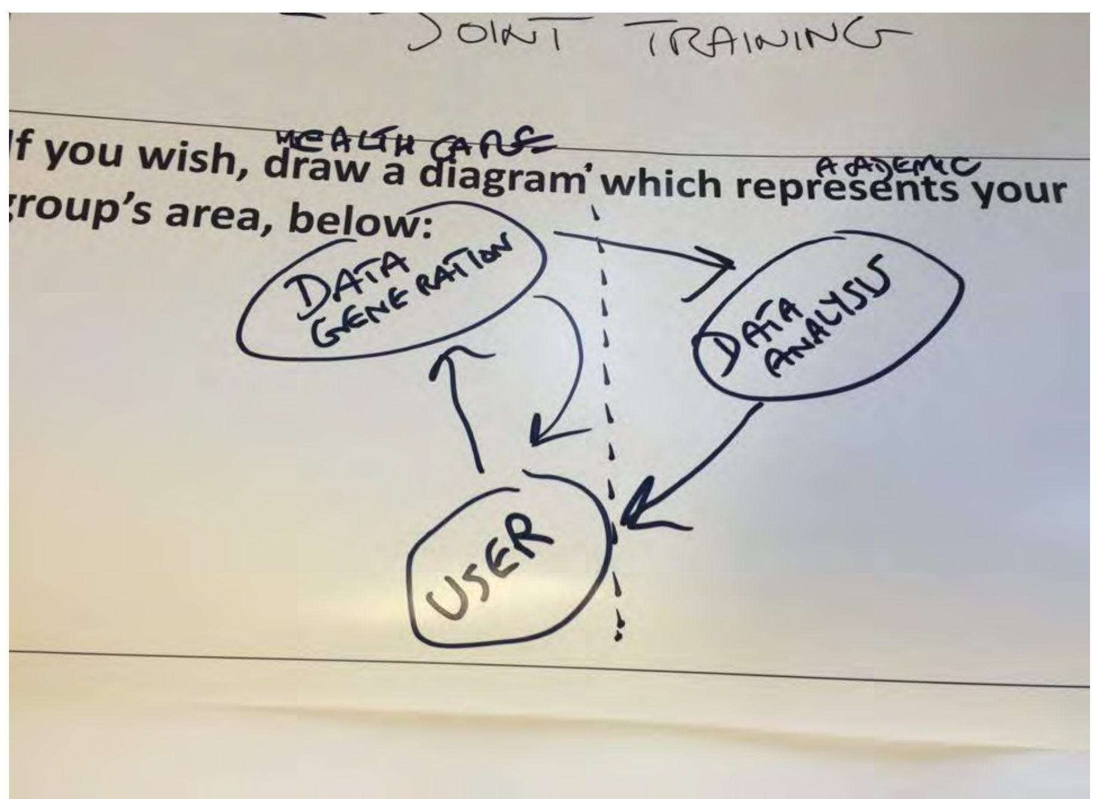
Broader Issue #2: existing & new (causal inference)

Do we need new methods before we train?
Funding focus, opportunities & limits? need revision
Data science method development & evaluation is a new grand challenge
Case studies and exemplars of the field in action

Broader Issue #3:Environment as a disincentive

- A regulatory? – not uniform or universal
- Mobility requirements/ professional incentive to leave one profession
- Lack of multi-disciplinarity funding for truly multidisciplinary programmes
- Joint training

If you wish, draw a diagram which represents your group’s area, below:



Group Members:
Tim Croudace
Jean Baptiste Cazier
Anne Cunningham
Mark Gilthorpe
Colin McGowan
Maxine Mackintosh

Group Name:	Colour: **Green 1**

Broader Issue #1:Data Quality

Cause:

- Instrumental errors, human errors, bias
- Time factor value of fields changes
- Ambiguity in meaning of data
- Represtativeness of samples

Broader Issue #2:Ethical + confidentiality: Can we use synthetic data?

Can we use pseudonymised/ anonymised/de-identified data?
Can we use data from other countries? (and map to UK context)
Getting consent in advance to donate data (card)

Broader Issue #3: Data semantics

- Meaning of labels change over time/datasets
- Add classes to data and then causal reasoning using semantic tags
- Add semantics/ontology labels

If you wish, draw a diagram which represents your group's area, below:	Group Members: **Eric Atwell** Georgios Aivaliotis Joao Bettencourt-Silva Samantha Crossfield Rajendra Kadel Daniel Neagu

Group Name:

Colour: **Green 2**

Broader Issue #1:Data quality

- Understanding of data collection
- Data sharing
- Need to record metadata

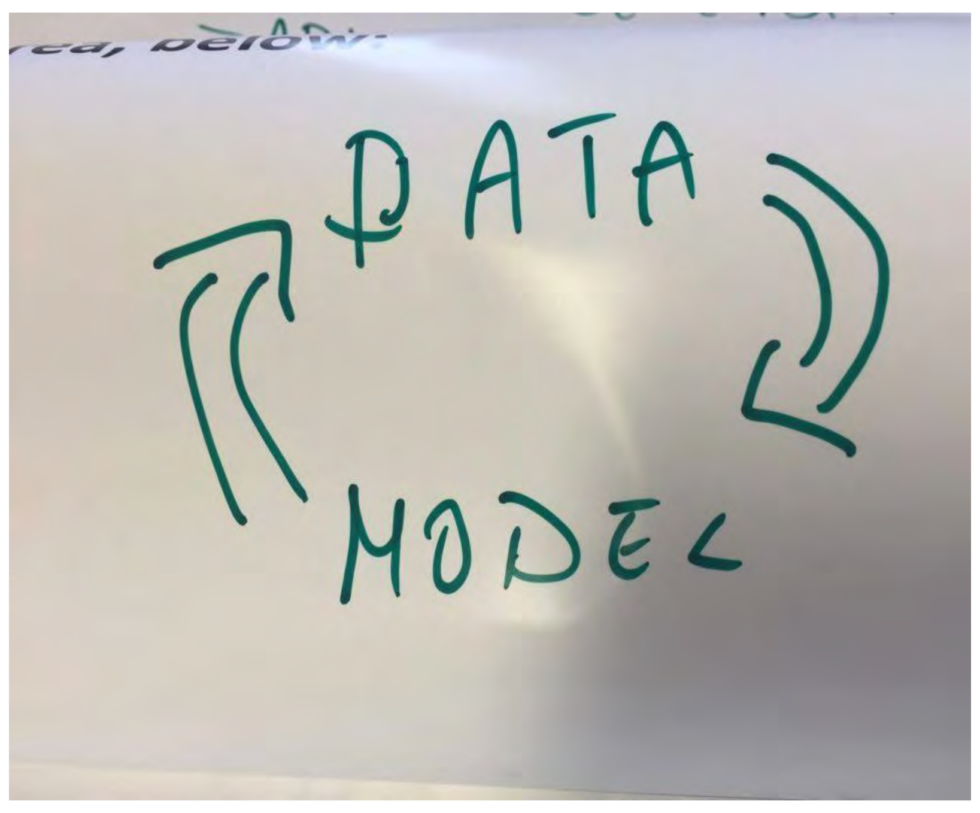
Broader Issue #2:Methods

- Causal dependencies (not just correlation)
- Artificial intelligence

Broader Issue #3:Combining data

Methods of data integration
Understanding context

If you wish, draw a diagram which represents your group's area, below:



Group Members:
Ji Ni
Vincenzi Nicosia
Iker Perez
Jenny Barrett
Edmore Champiwa

Group Name:

Colour: **Blue**

Broader Issue #1: Trustworthiness

- Develop a structure/model & criteria for 'what trustworthiness looks like' & demonstrate it – appetite for research proposals
- Changing to a culture of pulling data use rather than resisting it

-> civic uses->imperative->Ethical to use; not not use

- Context of data- places/uses etc
- What is different/taboo about health data?

Broader Issue #2: Control of data & linkage

- 'Ownership concept': is it helpful concept or does it confuse the conversation?
- Consent processes: opt-in/opt-out models etc
- Uses- duties, governance
- Who accesses/ controls data?
- Coproduction of care – different users, responsibility, purposes

Technical solutions

- Auditing access= consequences of breaching trust

Broader Issue #3: Expectations for use of data for public benefit + personal benefit

- What do you think should be/ is done + approve of?
- How much care is taken?

How do expectations for health data compare to other data collection? Eg Google, Tesco

- Literacy around benefits of data sharing is organisations/publics changing culture
- How much does it matter that you perceive the benefit now?

If you wish, draw a diagram which represents your group's area, below:

Group Members:

Ruth Norris Jackie Cassell
Mike Chantler Wendy Moncur
Jon Fistein Yousef Amar
David Hogg David Osler
Chris Taylor