# UK Health Data Analytics Workshop: Well Sorted Materials

**7th January 2016**

## Contents



**For an online, interactive version of the visualisations in this document, go here:**

**www.well-sorted.org/output/UKHDANTechnologyChallenges**

# Introduction

Dear participant,

Thank you for taking part in submitting and sorting your ideas.

This document contains several visualisations of your ideas, grouped by the average of your online sorts. They are:

Dendrogram - This tree shows each submitted idea and its similarity to the others. The lower two ideas 'join' the more people grouped those two ideas together. For example, if two ideas join at the bottom, every person grouped those two together.

Tree Map - This visualisation presents an 'average' grouping. It is calculated by 'cutting' the Dendrogram at the dashed line so that any items which join lower than that line are placed in the same group. In addition, rectangles which share a side of the same length are more similar to each other than their peers.

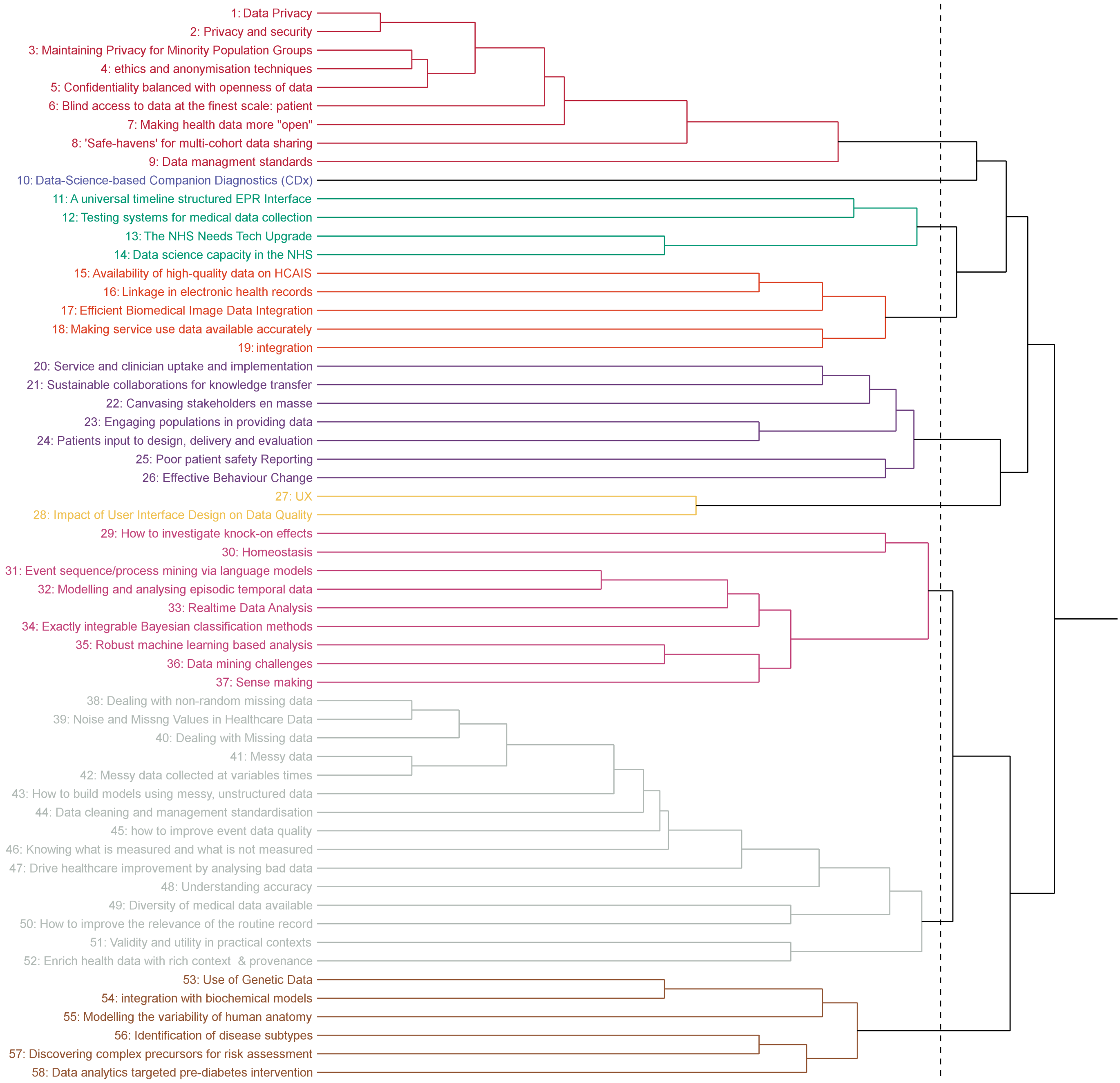
Heat Map - This visualisation shows a similarity matrix where each idea is given a colour at the intersection with another idea, showing how similar the two are. This is useful to see how well formed a group is. The more red there is in a group (shown by the black lines), the more similar the ideas inside it were judged to be.

Raw Group Data - This table shows every submitted idea and its longer description. They are shown in the same order as the Dendrogram (so similar ideas are close to each other) and split into the coloured groups used in the Tree Map. In addition, each idea has been given a unique number so they are easier to find.

# References

[1] Methven, T. S., Padilla, S., Corne, D. W., & Chantler, M. J. (2014, February). Research Strategy Generation: Avoiding Academic 'Animal Farm'. In Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work & social computing (pp. 25-28). ACM. doi>10.1145/2556420.2556785

# Dendrogram


1: Data Privacy
2: Privacy and security
3: Maintaining Privacy for Minority Population Groups
4: ethics and anonymisation techniques
5: Confidentiality balanced with openness of data
6: Blind access to data at the finest scale: patient
7: Making health data more "open"
8: 'Safe-havens' for multi-cohort data sharing
9: Data managment standards
10: Data-Science-based Companion Diagnostics (CDx)
11: A universal timeline structured EPR Interface
12: Testing systems for medical data collection
13: The NHS Needs Tech Upgrade
14: Data science capacity in the NHS
15: Availability of high-quality data on HCAIS
16: Linkage in electronic health records
17: Efficient Biomedical Image Data Integration
18: Making service use data available accurately
19: integration
20: Service and clinician uptake and implementation
21: Sustainable collaborations for knowledge transfer
22: Canvasing stakeholders en masse
23: Engaging populations in providing data
24: Patients input to design, delivery and evaluation
25: Poor patient safety Reporting
26: Effective Behaviour Change
27: UX
28: Impact of User Interface Design on Data Quality
29: How to investigate knock-on effects
30: Homeostasis
31: Event sequence/process mining via language models
32: Modelling and analysing episodic temporal data
33: Realtime Data Analysis
34: Exactly integrable Bayesian classification methods
35: Robust machine learning based analysis
36: Data mining challenges
37: Sense making
38: Dealing with non-random missing data
39: Noise and Missng Values in Healthcare Data
40: Dealing with Missing data
41: Messy data
42: Messy data collected at variables times
43: How to build models using messy, unstructured data
44: Data cleaning and management standardisation
45: how to improve event data quality
46: Knowing what is measured and what is not measured
47: Drive healthcare improvement by analysing bad data
48: Understanding accuracy
49: Diversity of medical data available
50: How to improve the relevance of the routine record
51: Validity and utility in practical contexts
52: Enrich health data with rich context & provenance
53: Use of Genetic Data
54: integration with biochemical models
55: Modelling the variability of human anatomy
56: Identification of disease subtypes
57: Discovering complex precursors for risk assessment
58: Data analytics targeted pre-diabetes intervention

# Tree Map

### RED

1: Data Privacy

2: Privacy and security

3: Maintaining Privacy for Minority Population Groups

4: ethics and anonymisation techniques

5: Confidentiality balanced with openness of data

6: Blind access to data at the finest scale: patient

7: Making health data more "open"

8: 'Safe-havens' for multi-cohort data sharing

9: Data managment standards

### BLUE

10: Data-Science-based Companion Diagnostics (CDx)

### GREEN

11: A universal timeline structured EPR Interface

12: Testing systems for medical data collection

13: The NHS Needs Tech Upgrade

14: Data science capacity in the NHS

### ORANGE

15: Availability of high-quality data on HCAIS

16: Linkage in electronic health records

17: Efficient Biomedical Image Data Integration

18: Making service use data available accurately

19: integration

### PURPLE

20: Service and clinician uptake and implementation

21: Sustainable collaborations for knowledge transfer

22: Canvasing stakeholders en masse

23: Engaging populations in providing data

24: Patients input to design, delivery and evaluation

25: Poor patient safety Reporting

26: Effective Behaviour Change

### YELLOW

27: UX

28: Impact of User Interface Design on Data Quality

### PINK

29: How to investigate knock-on effects

30: Homeostasis

31: Event sequence/process mining via language models

32: Modelling and analysing episodic temporal data

33: Realtime Data Analysis

34: Exactly integrable Bayesian classification methods

35: Robust machine learning based analysis

36: Data mining challenges

37: Sense making

### SILVER

38: Dealing with non-random missing data

39: Noise and Missng Values in Healthcare Data

40: Dealing with Missing data

41: Messy data

42: Messy data collected at variables times

43: How to build models using messy, unstructured data

44: Data cleaning and management standardisation

45: how to improve event data quality

46: Knowing what is measured and what is not measured

47: Drive healthcare improvement by analysing bad data

48: Understanding accuracy

49: Diversity of medical data available

50: How to improve the relevance of the routine record

51: Validity and utility in practical contexts

52: Enrich health data with rich context & provenance

### BROWN

53: Use of Genetic Data

54: integration with biochemical models

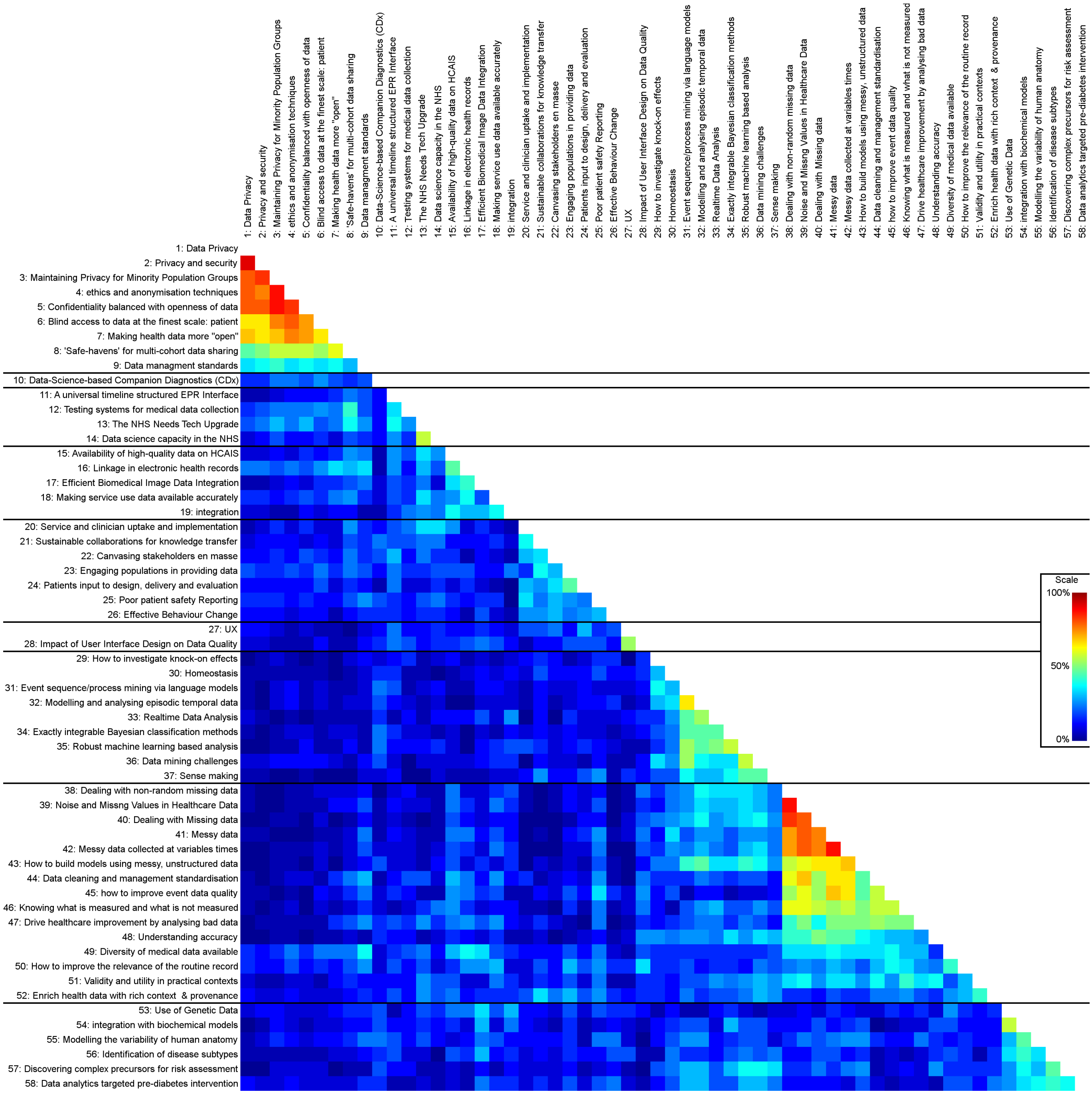
55: Modelling the variability of human anatomy

56: Identification of disease subtypes

57: Discovering complex precursors for risk assessment

58: Data analytics targeted pre-diabetes intervention

# Heat Map



1: Data Privacy
2: Privacy and security
3: Maintaining Privacy for Minority Population Groups
4: ethics and anonymisation techniques
5: Confidentiality balanced with openness of data
6: Blind access to data at the finest scale: patient
7: Making health data more "open"
8: 'Safe-havens' for multi-cohort data sharing
9: Data managment standards
10: Data-Science-based Companion Diagnostics (CDx)
11: A universal timeline structured EPR Interface
12: Testing systems for medical data collection
13: The NHS Needs Tech Upgrade
14: Data science capacity in the NHS
15: Availability of high-quality data on HCAIS
16: Linkage in electronic health records
17: Efficient Biomedical Image Data Integration
18: Making service use data available accurately
19: integration
20: Service and clinician uptake and implementation
21: Sustainable collaborations for knowledge transfer
22: Canvasing stakeholders en masse
23: Engaging populations in providing data
24: Patients input to design, delivery and evaluation
25: Poor patient safety Reporting
26: Effective Behaviour Change
27: UX
28: Impact of User Interface Design on Data Quality
29: How to investigate knock-on effects
30: Homeostasis
31: Event sequence/process mining via language models
32: Modelling and analysing episodic temporal data
33: Realtime Data Analysis
34: Exactly integrable Bayesian classification methods
35: Robust machine learning based analysis
36: Data mining challenges
37: Sense making
38: Dealing with non-random missing data
39: Noise and Missng Values in Healthcare Data
40: Dealing with Missing data
41: Messy data
42: Messy data collected at variables times
43: How to build models using messy, unstructured data
44: Data cleaning and management standardisation
45: how to improve event data quality
46: Knowing what is measured and what is not measured
47: Drive healthcare improvement by analysing bad data
48: Understanding accuracy
49: Diversity of medical data available
50: How to improve the relevance of the routine record
51: Validity and utility in practical contexts
52: Enrich health data with rich context & provenance
53: Use of Genetic Data
54: integration with biochemical models
55: Modelling the variability of human anatomy
56: Identification of disease subtypes
57: Discovering complex precursors for risk assessment
58: Data analytics targeted pre-diabetes intervention
Scale
100%
50%
0%

## Raw Group Data

| Colour | # | Title | Description |
|---|---|---|---|
| Red | 1 | Data Privacy | Healthcare has particularly critical privacy requirements. How can data privacy be guaranteed, monitored and tracked through a data science ecosystem in a way that is personalized to a user? |
| | 2 | Privacy and security | Availability, ownership and usage of data for healthcare purposes come with both challenges of protecting privacy and assuring cyber security functionalities. |
| | 3 | Maintaining Privacy for Minority Population Groups | Analysing a large healthcare data set may not pose considerable risk of re-identification for individual research subjects, but for small minority groups and isolated communities, with a particularly rare disease, maintaining privacy is more challenging |
| | 4 | ethics and anonymisation techniques | It is very challenging to anonymise identities and have consent from individuals in the healthcare system. In my opinion, the qualitative data should always support the quantitative ones, and this makes the ethics considerations more complex. |
| | 5 | Confidentiality balanced with openness of data | For data science to fully exploit the potential of health data, the integration of many types of data is required. The problem with data integration is that it poses risks of confidentiality as more types of hetergenous data are combined. |
| | 6 | Blind access to data at the finest scale: patient | Due to privacy and security getting data at individual or houselhold level is not possible. Nonetheless secure servers architecture with interoperability may allow computations at this level and render results at a higher geographical scale. |
| | 7 | Making health data more "open" | We are plagued by data sharing barriers: governance/disclosure restrictions, intellectual property issues or unsustainable to move large datasets. Requires an ethical-legal-social-technical solution to make biomedical data more accessible and reusable. |
| | 8 | 'Safe-havens' for multi-cohort data sharing | There are a plethora of existing healthcare and social data in the UK but the single biggest barrier to effective linkage is governance. A pragmatic solution is the use of data safe-havens (e.g. Farr Institutes). |
| | 9 | Data managment standards | As a researcher involved in the experimental biomedical data management and analysis (RCUK grants) I found a lack of well characterized standards as one of the obstacle that effect my work. |

| Colour | # | Title | Description |
| --- | --- | --- | --- |
| **Blue** | **10** | Data-Science-based Companion Diagnostics (CDx) | Data Science can personalise healthcare. Drug regulators (FDA/EMA) may approve CDx to limit treatment only to those likely to benefit or not be harmed. Data-Science-based CDx must be strictly validated, available everywhere, and unchanged for years. |

| Colour | # | Title | Description |
| --- | --- | --- | --- |
| **Green** | **11** | A universal timeline structured EPR Interface | The timeline layered graphically rich EPR (cf UHS-Lifelines) offers a universal data structure for the integration & presentation of primary, 2ndary & social care records. The challenge is to implement this powerful NHS IT tool at local & national level. |
| | **12** | Testing systems for medical data collection | Although the miniaturisation of technology opens new possibilities of collecting healthcare data via wearable devices, testing prototype devices with participants suffering from neurological conditions often involves a lot of paperwork, delaying research. |
| | **13** | The NHS Needs Tech Upgrade | The NHS should take advantage of recent technological advancements in record keeping and real-time data collection as an opportunity to collaborate with researchers in utilising this info to help patients. Issues: confidentiality? Cost? |
| | **14** | Data science capacity in the NHS | Developing the skills across the NHS to understand what data science is and why it matters |

| Colour | # | Title | Description |
|---|---|---|---|
| Orange | **15** | Availability of high-quality data on HCAIS | It has been shown that integrating different hospital data sources can yield enhanced information for surveillance of a wide range of threats and this illustrates the need for integration of healthcare data from all sources. |
| | **16** | Linkage in electronic health records | The lack of direct linkage between some primary and secondary care databases can lead to difficulties in successfully tracking patients across centres. However, implementing a fully integrated system in the UK across all healthcare sites is problematic. |
| | **17** | Efficient Biomedical Image Data Integration | What kind of infrastructure enables data analytics involving very large, distributed clinical (PACS) and Life Science image data repositories, combining cloud computing, image compression and semantic web technologies? A multi-disciplinary challenge. |
| | **18** | Making service use data available accurately | Data on usage and access of health services is not properly recorded, nor widely available for researchers. Proposals to link health service use data to survey data have been made but how these will be carried out has not been forthcoming |
| | **19** | integration | how to integrate data from wearable sensors with information from traditional healthcare systems. |

| Colour | # | Title | Description |
|---|---|---|---|
| Purple | 20 | Service and clinician uptake and implementation | Clinicians often act as gatekeepers of mental health services. Therefore, clinicians views and attitudes effect type of healthcare received. Issues about how services manage complexities of real time, in-the-moment data collection. |
| | 21 | Sustainable collaborations for knowledge transfer | Sustainable collaboration and effective knowledge transfer will require the development of infrastructures and the formation of teams that share a common understanding of the potential benefits of data science for health care as well as its limitations. |
| | 22 | Canvasing stakeholders en masse | Extending tools like Well-Sorted and open innovation systems to harness the power of massed stakeholders to provide input on for instance, best practice patient support. |
| | 23 | Engaging populations in providing data | Much health care data is routine data on utilisation and costs, and lacks the other critical aspect for cost-effectiveness - quality of life and patient experience. How do we develop methods to collect patient reported data at scale? |
| | 24 | Patients input to design, delivery and evaluation | Need to make the data science tools accessible so that patients, professionals, managers, and policy folk can engage in the design, delivery and evaluation of healthcare. Research anticipated benefits and risks - and also to discover unintended outcomes. |
| | 25 | Poor patient safety Reporting | When it comes to patient safety data and incidents, there is a culture of poor and under reporting. Decisions made based on existing data tend to lack credibility due to poor confidence in the available data. |
| | 26 | Effective Behaviour Change | Using on line technology to change people's behaviour to improve wellbeing, follow treatment plans, and change lifestyle. Linking data analytic outcomes to low cost (ICT mediated) behaviour change. |

| Colour | # | Title | Description |
|---|---|---|---|
| Yellow | 27 | UX | Understanding the user-facing aspects required to convey complex information to multiple stakeholders. |
| | 28 | Impact of User Interface Design on Data Quality | Data quality issues resulting from variations in user interface design are non random and can act as a confounding factor on data science based on routine data. The causal relationship is evident but poorly understood. |

| Colour | # | Title | Description |
|---|---|---|---|
| Pink | **29** | How to investigate knock-on effects | Health data is often analysed in a series of discrete pipeline stages, but users rarely investigate the sensitivity of findings to decisions made earlier on during analysis. How can data science increase rigour by joining together this 'broken' workflow? |
| | **30** | Homeostasis | A characteristic of healthcare data is the fact that the time-series arise from a system with strong homeostasis - it is an individual that is actively attempting to restore itself to normality - this is a unique constraint for data science applications. |
| | **31** | Event sequence/process mining via language models | Standard data mining classifies and clusters unordered sets of data. health data has ordered sequences of events, more suited to language models from linguistics, e.g. n-gram taggers, Brill taggers and Chart parsers for tagging Part-of-Speech sequences. |
| | **32** | Modelling and analysing episodic temporal data | Consider recognising significant change in an individual's behaviour from wearable/IoT data. Since every day is different the challenge is how to go beyond crude averages (eg time spent sleeping, walking etc) to learn what is normal for that person. |
| | **33** | Realtime Data Analysis | There are many frameworks available for processing streaming realtime data (from medical sensors). Are these widely used? Can they be adapted to work with the tools that researchers currently use? Discuss. |
| | **34** | Exactly integrable Bayesian classification methods | Extracting clinically predictions from genomic data is presently done via so-called `gene signatures', which are poor man's alternatives to proper regression. One would prefer Bayesian methods, but in high dimensions they pose prohibitive CPU demands. |
| | **35** | Robust machine learning based analysis | The integration of machine learning approaches with traditional statistics is essential to deal with non-linearity and related variables within big data. A major associated challenge is the education of healthcare professionals in this approach. |
| | **36** | Data mining challenges | Data mining can unintentionally be misused, and can then produce results which appear to be significant; but which do not actually predict future behavior and cannot be reproduced on a new sample of data and bear little use. |
| | **37** | Sense making | Connecting analytical results with domain knowledge! |

| Colour | # | Title | Description |
|---|---|---|---|
| Silver | **38** | Dealing with non-random missing data | Data can sometimes be included or excluded due to it's relavence to an outcome, such as GP's being more likely to update or record smoking or BMI data if they think it is likely to become a health issue. This can make account for these measures difficult |
| | **39** | Noise and Missng Values in Healthcare Data | Healthcare data may have noise and missing values. If these are not handled appropriately, the results of the analysis may not be accurate. Human knowledge and data pre-processing methods can be used to clean noise and handle missing values in data. |
| | **40** | Dealing with Missing data | Missing data can be informative and hence need state-of-the-art modelling techniques to overcome this problem. |
| | **41** | Messy data | Dealing with missing and low quality data - managing uncertainty and interpolating between data points. Identifying which points are more likely to be accurate. |
| | **42** | Messy data collected at variables times | Routinely collected are messy (not systematically recorded) and collected at different points in time |
| | **43** | How to build models using messy, unstructured data | Statistical modelling is based on samples of complete data arising from a designed experiment. But healthcare data has a complex data generating process, missingness and lack of structure. |
| | **44** | Data cleaning and management standardisation | Analysing data for health care research normally requires a stage of cleaning and subsequent manipulation before analysis can be undertaken. This stage is often done by the researcher rather than at data source introducing repetition and inconsistency. |
| | **45** | how to improve event data quality | how do we improve the quality of the data to underpin information and knowledge in delivering healthcare |
| | **46** | Knowing what is measured and what is not measured | One of the biggest challenges of using routine health data is knowing what gets measured, when, by whom and to what standard. Routine NHS systems are not often equipped for the purposes of large scale data extraction which limits research potential. |
| | **47** | Drive healthcare improvement by analysing bad data | NHS IT systems are a mess. Operational data is a mess, and many patient records have major errors. Many systems use proprietary "standards". Data can take weeks to flow end to end (even if it gets there) ... the mess needs exposing and critiquing! |
| | **48** | Understanding accuracy | We're all familiar with papers where the algorithm performance results are optimistic (to put it politely). Is this just bad practice or do we need new tools to understand the performance of algorithms when working with Big Data? |
| | **49** | Diversity of medical data available | Medical records fall in the category of 'big data'. However, healthcare data is also multi-modal and extremely complex and noisy. A huge data science research challenge is to find methods to incorporate data of different types and from different sources. |

| | | | |
|---|---|---|---|
| | **50** | How to improve the relevance of the routine record | Methods are needed to address the inadequate attention paid to outcome measurement in clinical care, research, and consumer health applications, in order to reduce avoidable waste in health data analytics. |
| | **51** | Validity and utility in practical contexts | Tools and methods developed must have validity and utility for clinicians and medical professionals in practical contexts. Factors such as technical infrastructure, required expertise, data quality and usage context must be considered in their design. |
| | **52** | Enrich health data with rich context & provenance | Representation of data at the point of care is disappointing. For trrue Learning Health systems clinicians need rich context/provenance, as well as the ability to synthesise new best practice through system learning. Secondary use data isn't sufficient. |

| Colour | # | Title | Description |
|---|---|---|---|
| **Brown** | **53** | Use of Genetic Data | New methods are needed to store, access and interpret genetic data and its impact on human health. |
| | **54** | integration with biochemical models | Making genetic or protein databases with more knowledge of function in the cell, of facilitate hypothesis formation and reduce false positives on data mining. |
| | **55** | Modelling the variability of human anatomy | There are millions of images of parts of human bodies in hospital databases. It should be possible to use them to build statistical models of how human anatomy varies across the population. However, the data is very variable and only weakly annotated. |
| | **56** | Identification of disease subtypes | Large data sets if linked data sets allow us to identify disease subtypes that are currently not captures by existing disease categories (e.g. Cancer grade). The identification of these and the associated statistics pose significant challenges. |
| | **57** | Discovering complex precursors for risk assessment | Opportunity: simultaneous emergence of large clinical bioresources (eg UKBB), distr'd powerful computing platforms, analytics algorithms, & medical signal/img analysis. Challenge: discover complex, heterogeneous sets of 'measures' predicting disease risk. |
| | **58** | Data analytics targeted pre-diabetes intervention | Data analytics can be used to predict which people are more likely to develop diabetes, based on various lifestyle markers, and anti-diabetic medication can be prescribed to this population. This is a goal of the Diabetes Prevention Programme. |