Dr Liz Ford and Dr Imogen Rogers

Brighton and Sussex Medical School

Using routinely collected data

Session Overview

- What do we mean by routinely collected health and care data?
- How is this different from other research data, and what additional skills and understanding do we need to use it?
- Under what legal basis can it be used and how do we get access?
- What types of datasets are available and how long does it take/how much does it cost?
- Can we link data to our research datasets?
- An example study we have carried out using linked data
- What resources are available to get me started and who can I ask for help?

What do we mean by routinely collected health and care data?

Data collected in the course of patient interactions with the health service

This could include:

Primary care:

- GP clinic notes (unstructured), prescriptions, and coded data (Read or Snomed Codes)

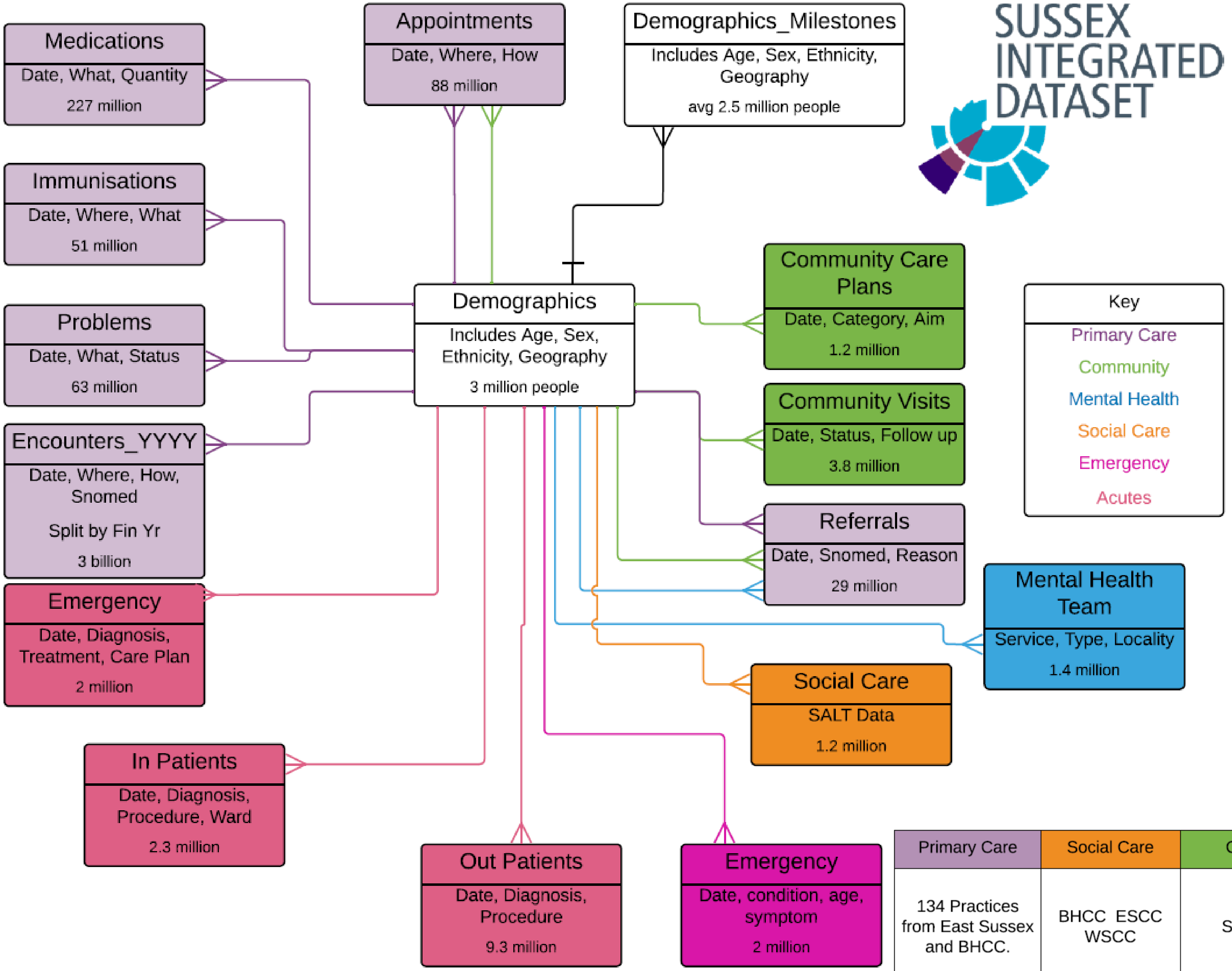
Secondary care:

- Hospital coded data – ICD codes – recorded for reimbursement for episodes of care:
    - Emergency Department
    - Outpatient Care
    - Inpatient Care
    - Discharge summaries (unstructured)
- Mental health patient notes (unstructured) and some coded data
- Nursing records

Disease Specific/Tertiary Care

- Disease registries e.g. cancer, MI, infectious diseases.

Social care data (on provision of SC by local authorities)



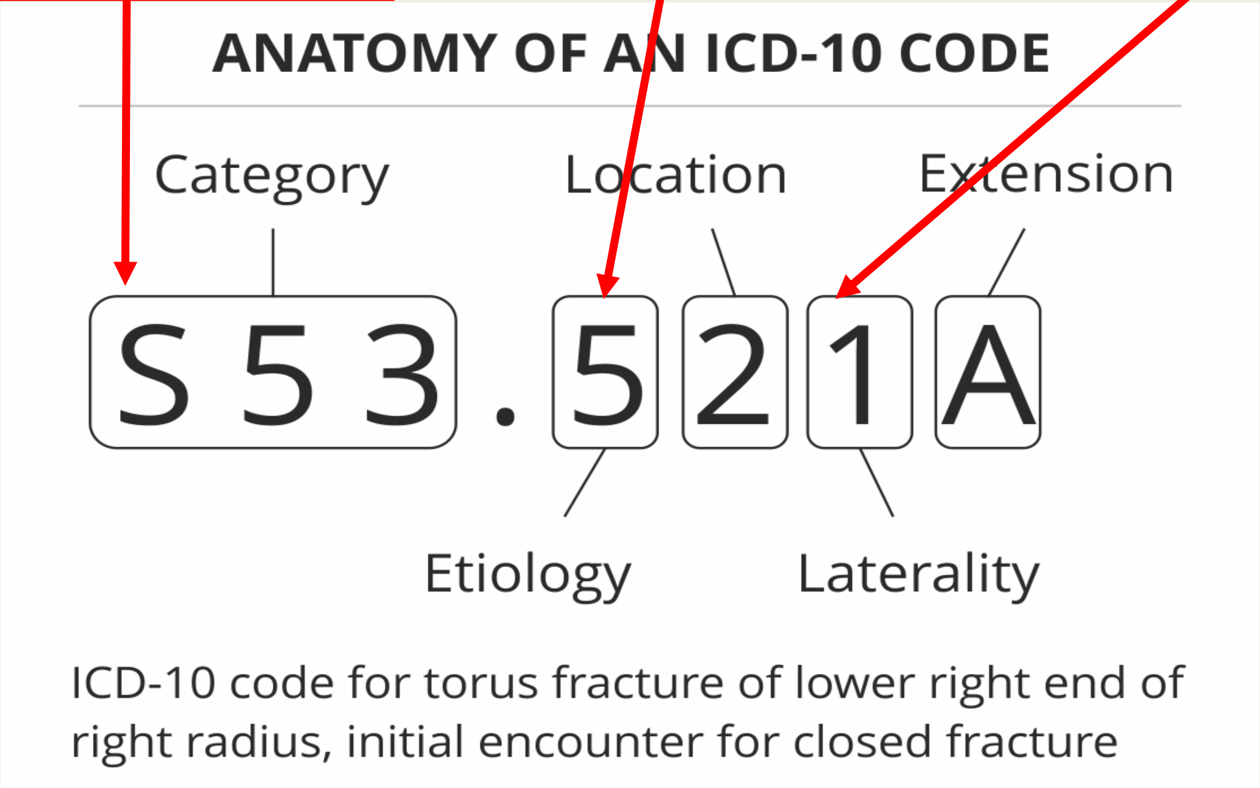
Primary Care	Social Care	Community	Mental Health	Acutes	Emergency
134 Practices from East Sussex and BHCC.	BHCC ESCC WSCC	SCFT ESHT	SPFT	UHSx, ESHT, QVH	SECAMB

Hospital Data – ICD10

.5 – Lower end of radius

S53 – Fracture of Forearm

.521 – 2 =Torus Fracture, 1 =Right Arm



ICD 10 code

S 8 6 • 0 1 1 D

Category	Etiology, Anatomic Site, Severity, Other Vital Details	Extension
S = Injuries, poisoning & certain other consequences of external causes related to single-body regions	S86.0 = Injury of Achilles tendon S86.01 = Strain of Achilles tendon S86.011 = Strain of right Achilles tendon	Primarily used to document episode of care for injuries & other conditions with external causes
S86 = Injury of muscle, fascia and tendon at lower leg		**A = Initial Encounter** **D = Subsequent Encounter** **S = Sequela**

https://icd.who.int/browse10/2019/en

What is coded in a GP record?

"Problem" / Symptoms

History / Family History

Test Results

Diagnosis?

Action Plan / Treatment / Referral

Read or SNOMED codes

FORD, Elizabeth (Dr) | Gender: Female | EMIS No: 15933 | Usual GP: HASSAN, Lamiece

Problems

16A2.00 Stiff Neck

13JM.13 Stress at work

1BT..11 Low Mood

Alcohol consumption	5 units/week
Body Mass Index	22.5
BP mmHg	127/88
Smoking Status	Never smoked

Known Allergies

14L1.00 H/O penicillin allergy
H170.11 Hay fever – pollens

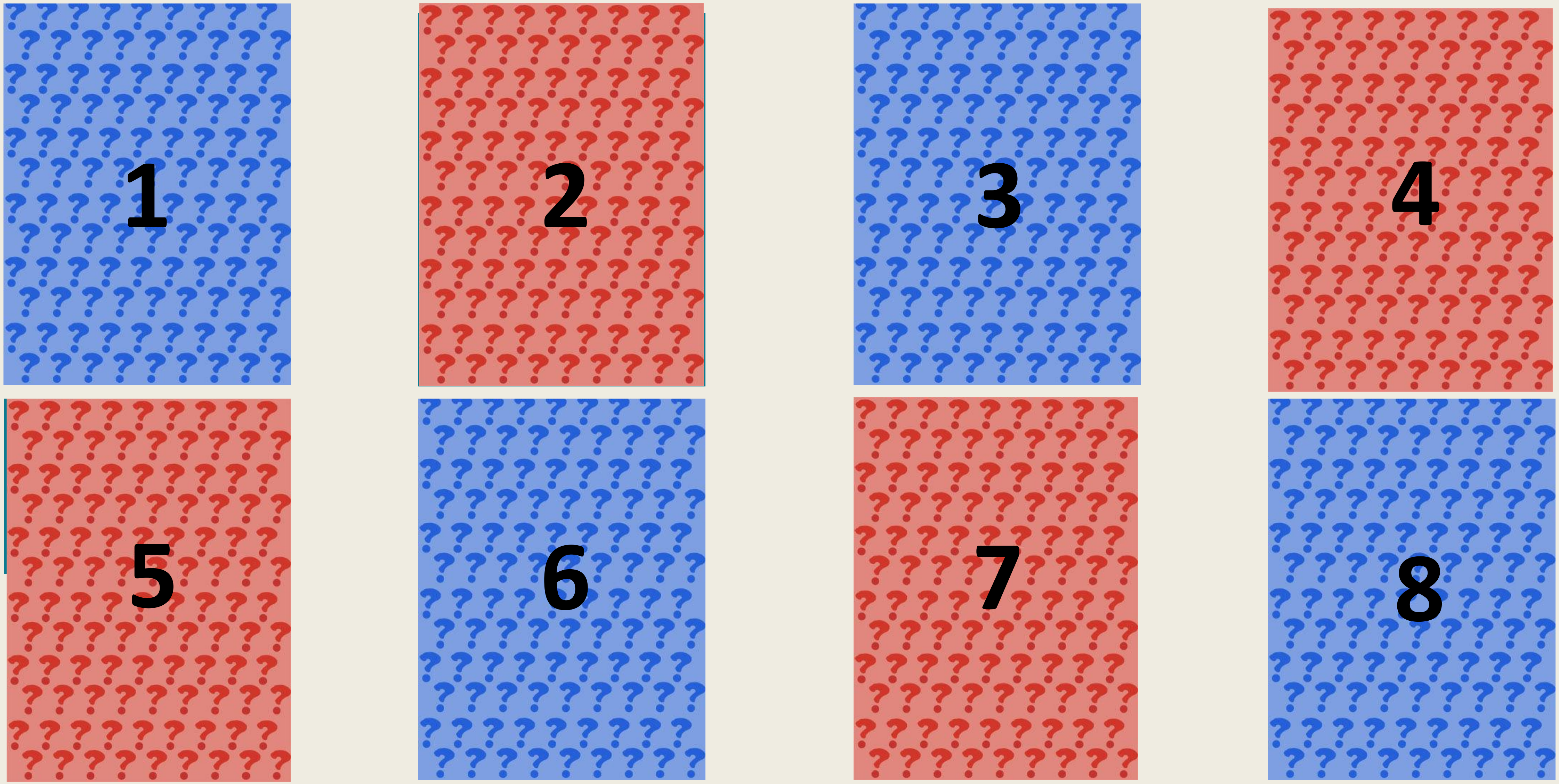
Current Consultation

Neck pain. Pain getting worse. Affecting driving, difficulty turning head. Patient reports no injuries.

Trouble sleeping, affecting mood. Patient distracted by work issues.

Medications

THE FAIRY TALE EDITION


1
2
3
4
5
6
7
8

Sleeping Beauty

brighton and sussex
medical school

Sleeping beauty

634..12	Female baby
13y8.00	Black Magic
TG3y500	Accident Caused by Spinning Machine
1BX1.00	Excessive Sleep
U12Ay00	Contact with plant thorns and spines and sharp leaves, occurrence at other specified place
S603.00	Concussion with more than 24 hours loss of consciousness …
8731.00	Manual Resuscitation
S603.00	… Return to Pre-existing Consciousness Level
1332.00	Married

Unstructured Data

- Also known as **free text**
- Needs a lot of pre-processing to structure it for statistical analysis, now using AI systems to pull out relevant clinical concepts (diagnoses, symptoms, medications). Branch of computer science called **Natural Language Processing (NLP)**
- For Sussex and Kent Mental Health data, AI is provided by Akrivia Health using the CRIS system. Apply for Sussex data via the SPFT CRIS Administrator: anne.watts@spft.nhs.uk asking for an application form. (Not sure on Kent details).
- De-identifying free text data is much harder than structured data, and privacy risks remain even if all identifiers are removed (social context, unique combinations of events or circumstances). Data custodians are VERY cautious about sharing text data.
- Akrivia structure (pseudo-code) the data before giving access to researchers

What does AI need to be able to "read"?

There are no signs of inflammation

He denies being a smoker

Mother has diabetes

She stopped taking her Sertraline last week

Symptoms suggest rheumatoid arthritis

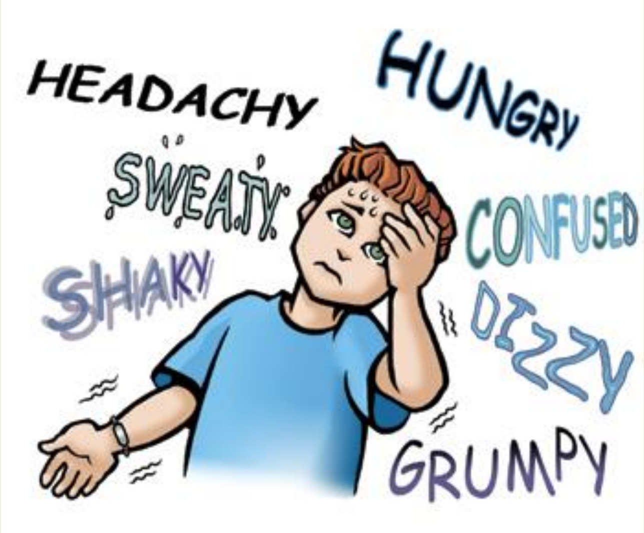
Negation

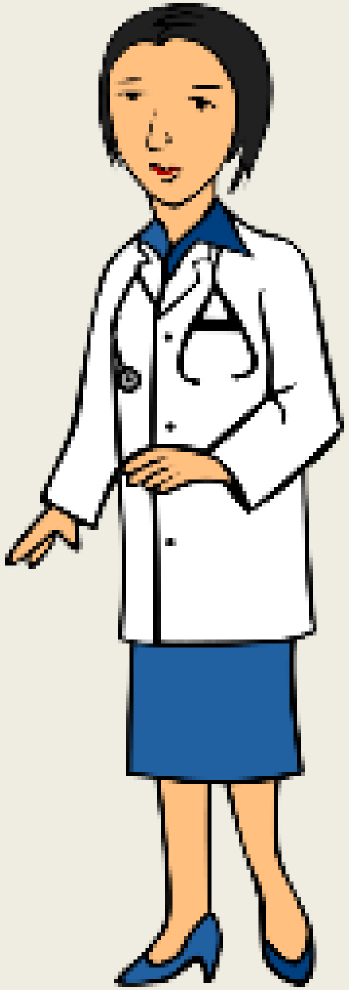
Experiencer

Time

Level of uncertainty

How is routine data different from other research data, and what additional skills and understanding do we need to use it?



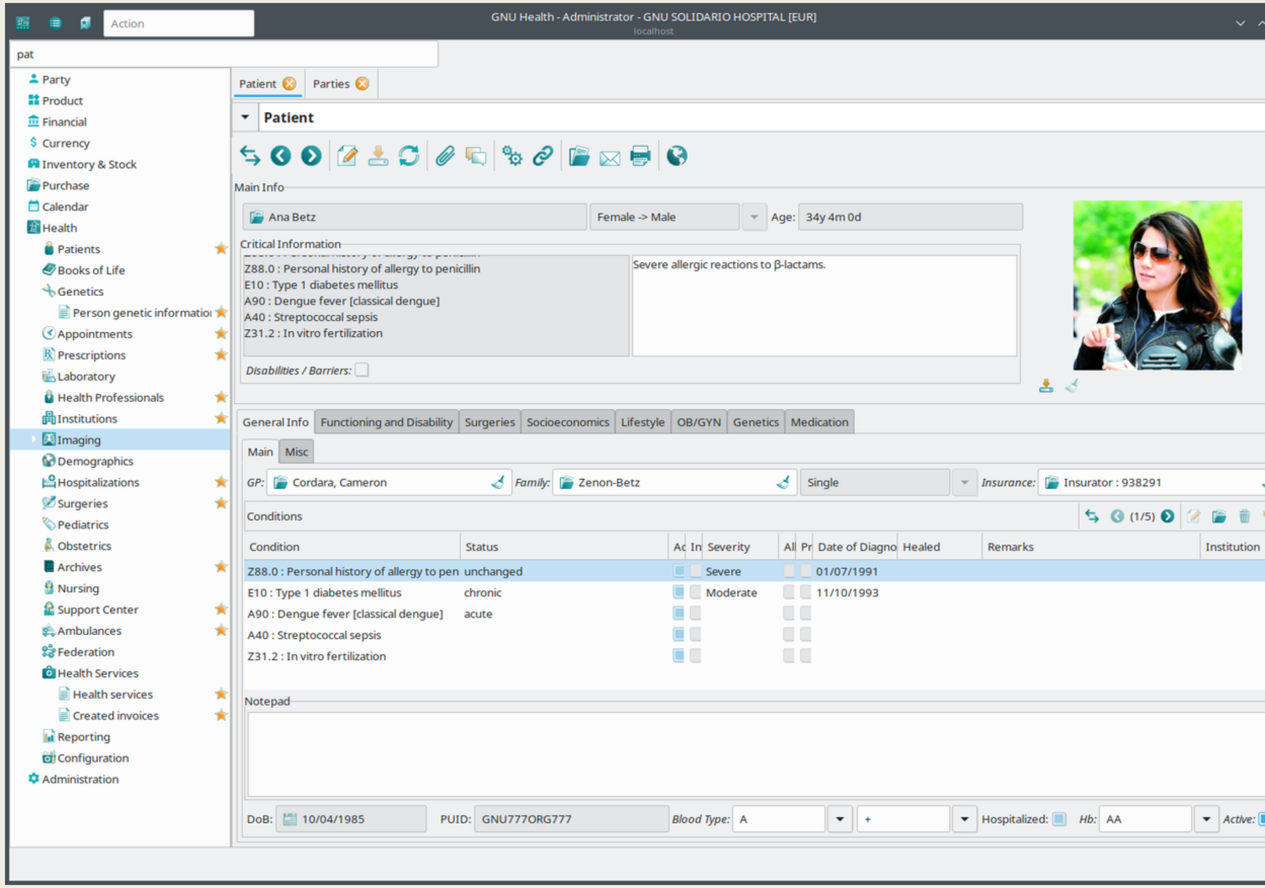


AKRIVIA HEALTH

brighton and sussex
medical school

Constraints on what data gets entered

- **Coding structure** – flexibility and limits
- **Patient record software interface** – ease of use
- **Time available** for documentation
- **Motivation for documentation** – billing / record for own memory / inform next clinician / medico-legal safety netting?
- **Clinical reasoning and filtering of information** – what is most important?
- **Consideration (or lack thereof) of secondary purposes**



How does data appear when data provided for research?

Usually – there is one row for each patient encounter with the system, with a date stamp, and a clinical code.

One patient could have thousands of rows.

How do you know if that patient should be in your study cohort?

Need to define: disease/condition, study period, exclusion criteria up front.

Disease/condition usually defined by a **list of codes** representing a firm diagnosis of the condition.

Table 4.
Example of NPPD patient tracking.

VISN	Station	Pt. ID	HCPC Code	L Code Description	
8	TAMPA/FL:	117748	L5300	BK, endo sys	17-Nov-98
8	TAMPA/FL:	117748	L5620	Test socket	17-Nov-98
8	TAMPA/FL:	117748	L5629	Acrylic socket	17-Nov-98
8	TAMPA/FL:	117748	L5637	Total contact	17-Nov-98
8	TAMPA/FL:	117748	L5662	Socket insert, silicone gel	17-Nov-98
8	TAMPA/FL:	117748	L5667	Suction suspen with locking pin	17-Nov-98
8	TAMPA/FL:	117748	L5910	Alignable system	17-Nov-98
8	TAMPA/FL:	117748	L5940	Ultra-light material	17-Nov-98
8	TAMPA/FL:	117748	L5962	Protective outer cover	17-Nov-98
8	TAMPA/FL:	117748	L5981	Flex walk system	17-Nov-98
9	LOUISVILLE/KY	117748	L5300	BK, endo sys	21-Jun-99
9	LOUISVILLE/KY	117748	L5620	Test socket	21-Jun-99
9	LOUISVILLE/KY	117748	L5629	Acrylic socket	21-Jun-99
9	LOUISVILLE/KY	117748	L5637	Total contact	21-Jun-99
9	LOUISVILLE/KY	117748	L5667	Suction suspen with locking pin	21-Jun-99
9	LOUISVILLE/KY	117748	L5669	Suction suspen w/o locking pin	21-Jun-99
9	LOUISVILLE/KY	117748	L5910	Alignable system	21-Jun-99
9	LOUISVILLE/KY	117748	L5940	Ultra-light material	21-Jun-99
9	LOUISVILLE/KY	117748	L5962	Flex walk system	21-Jun-99

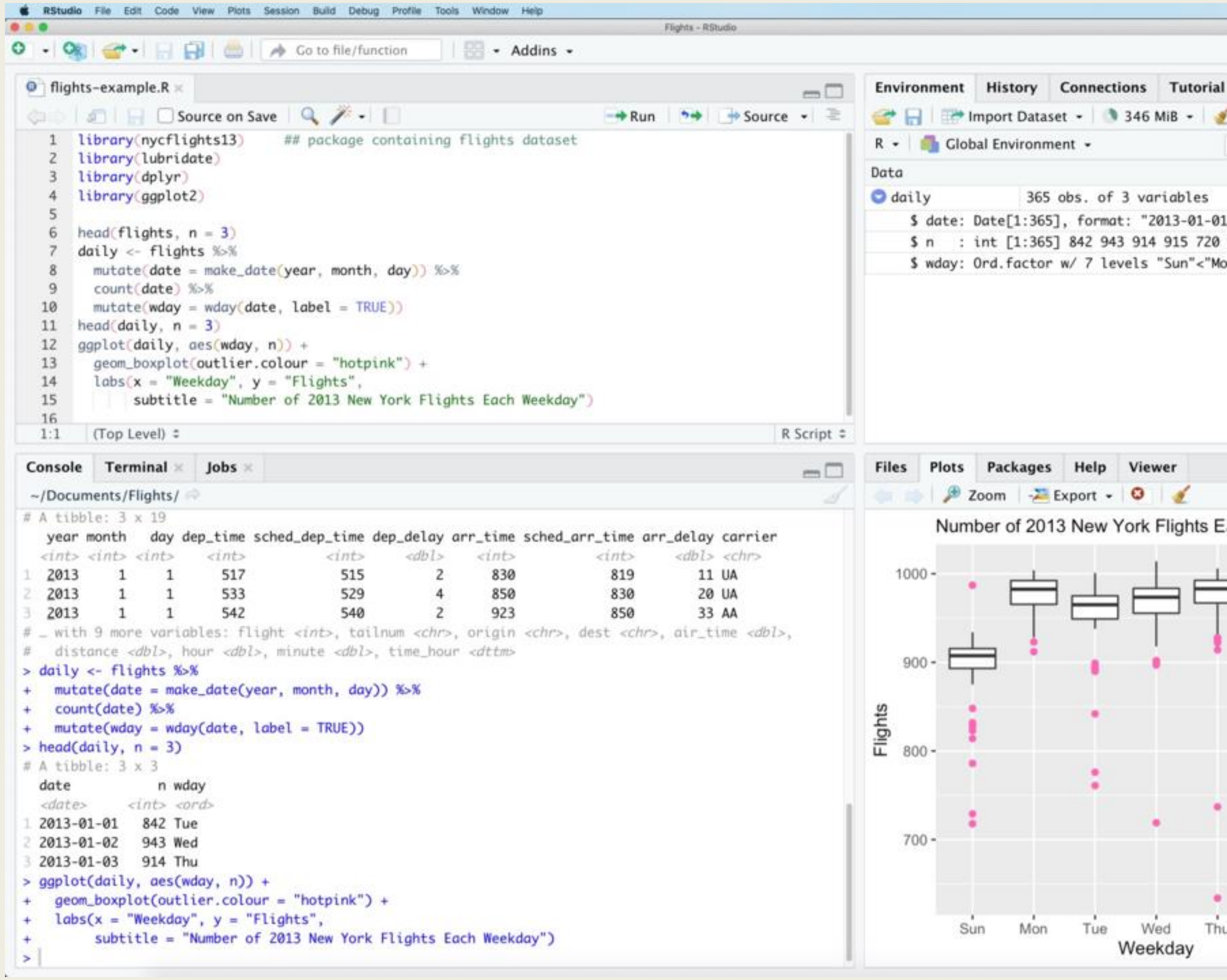
What statistical skills will I need?

Be able to **write code (syntax) in STATA, R or Python, with/without SQL**

Be prepared for lots of (months of) data processing, merging in code lists, checking cohorts, excluding outliers, using look-up tables etc.

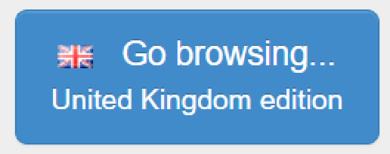
Understand how to reformat data (e.g. from long form to wide form) ready for analysis, keeping in dates where necessary e.g. earliest date of diagnostic code

Run analysis packages e.g. Cox regression and interpret outcomes

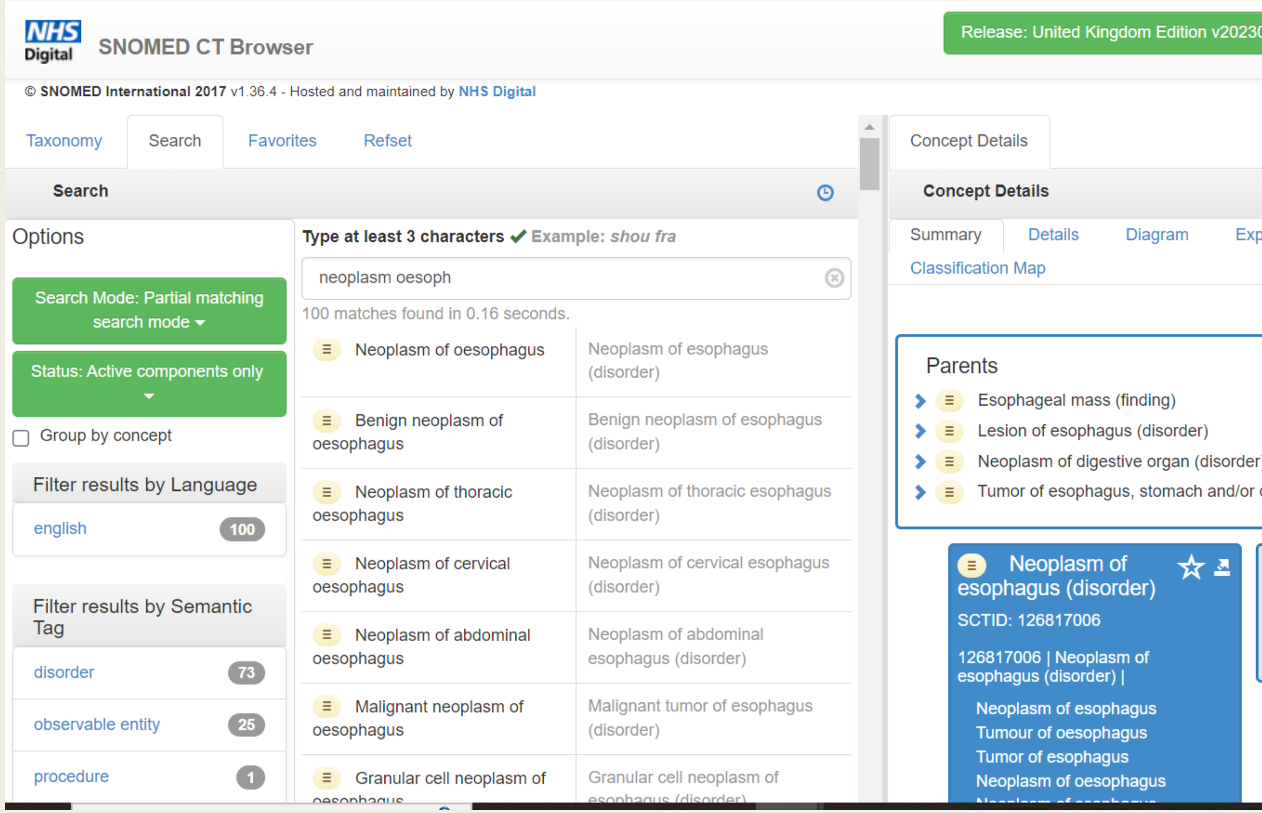


brighton and sussex medical school

Activity: Create a code list

- Put https://termbrowser.nhs.uk/? Into your web browser.
- Choose 

- Choose search from tabs on top left.
- Search for codes for diagnosis of:
  - Prostate cancer
  - Ovarian cancer
  - Oesophageal cancer
- Exclude any codes which don't represent diagnosis of the condition in the patient.
- Click on code to see **parent/child codes.**
- Count up how many codes you get and which represent the condition.
  - Remember synonyms esp. neoplasm, carcinoma
  - Use branching tree structure of codes to find more specific ones



5 Minutes!

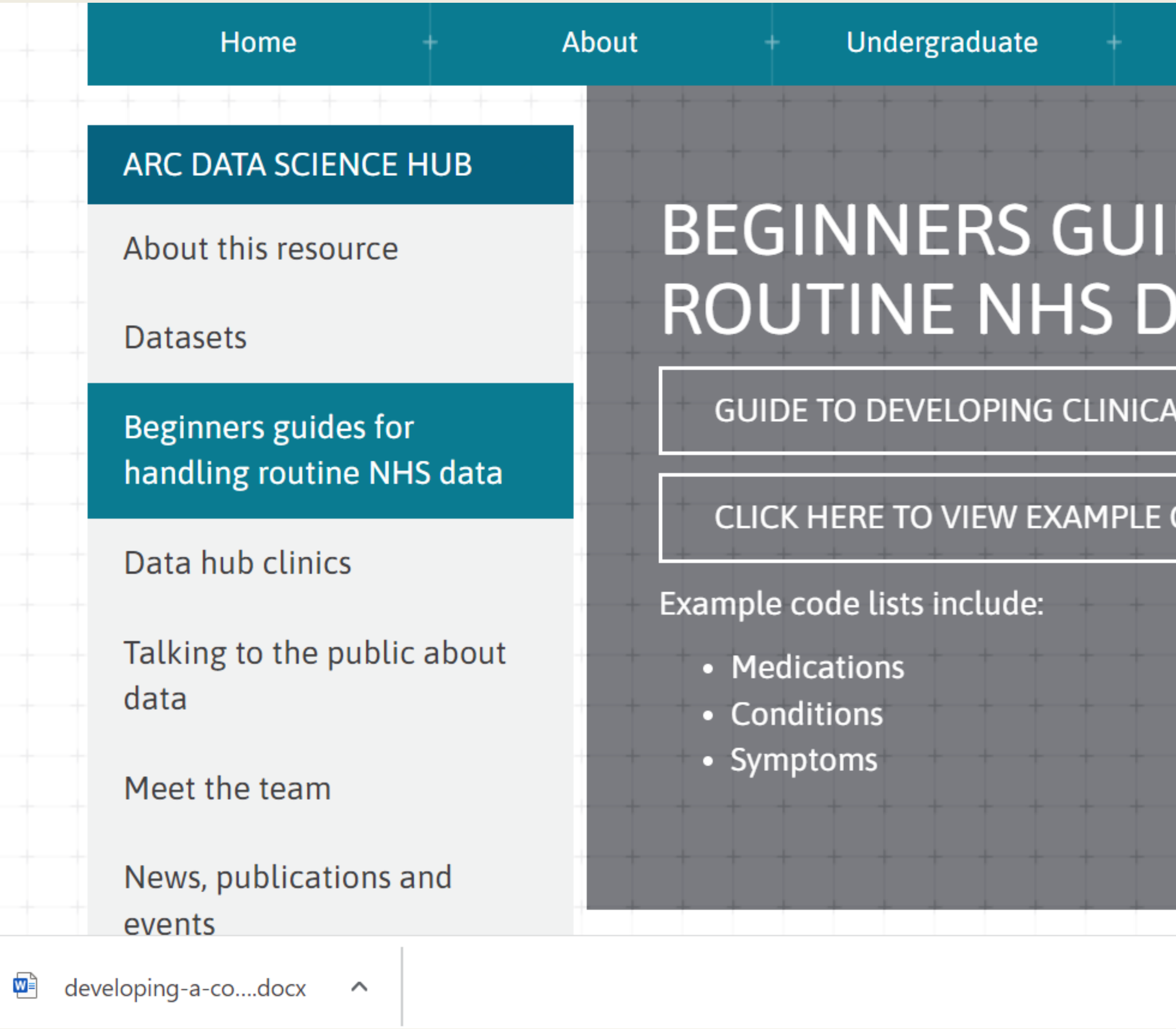
Type in chat: the condition you chose and the number of codes in your final list.

Conclusion…

Can you create a code list in 5 minutes? YES/NO

Tips:

- Always consult a clinician to get more guidance on what should be in and out of your code list.
- Use Imogen's code list resource to help think about the issues – www.bsms.ac.uk/arcdatahub
- Remember to think about other conditions of exclusion etc



Under what legal basis can routine data be used and how do we get access?

Lawful Basis of Processing Data

Data protection legislation requires that the collection and processing of personal data is fair, lawful and transparent. Lawful basis of use:

- ✓ Consent OR
- ✓ Task in the public interest
- ✓ with: Data minimisation (de-personalisation)

Processing of de-personalised patient data in universities, NHS organisations, or local authorities for research and service planning is usually done as a 'task in the public interest'.

(this is the legal basis for processing).

No patient should be identifiable from the data being used.

Why not just ask for consent?

There are several reasons why an **opt-out** approach may be preferable to **opt-in**:

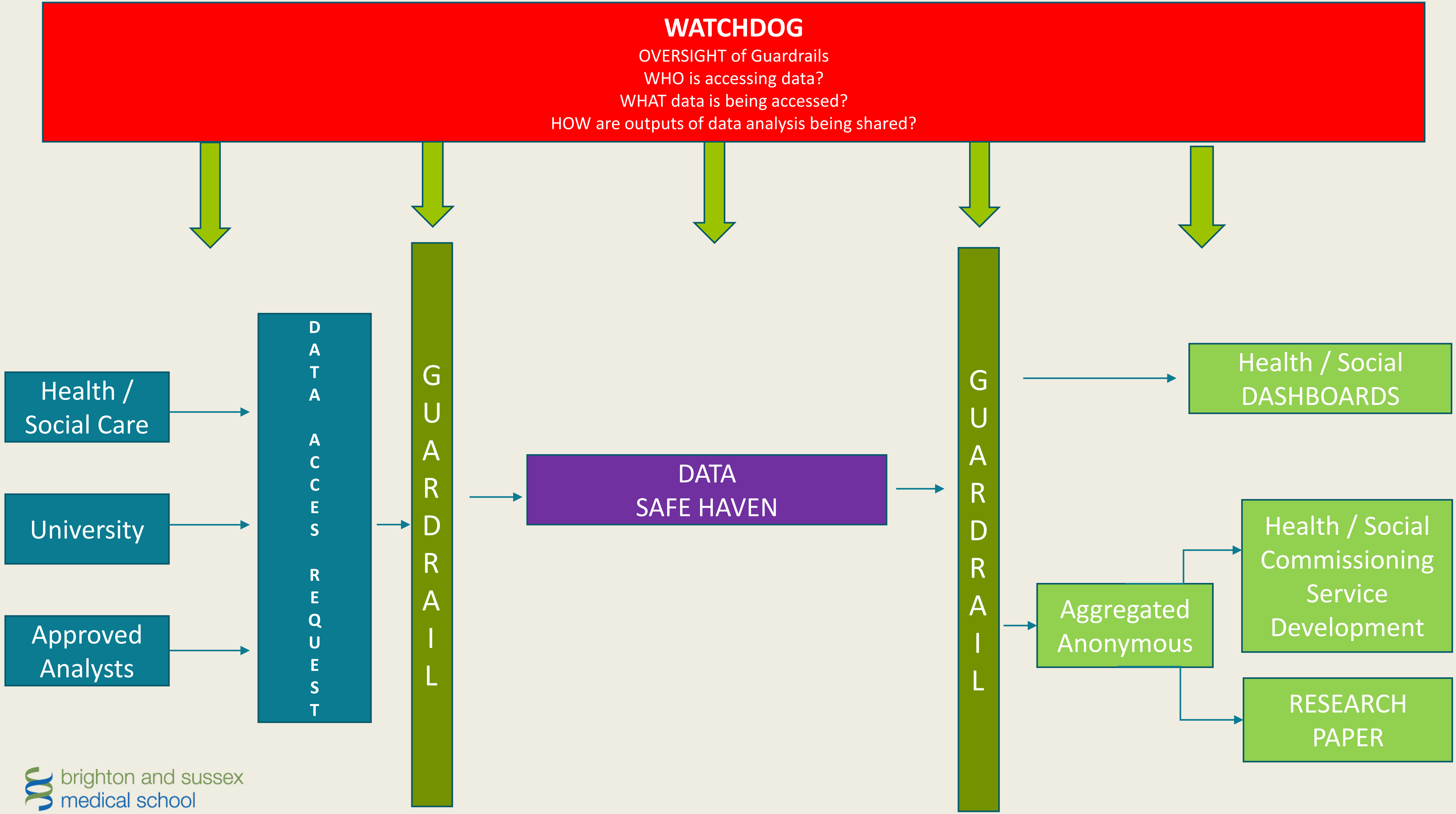
❖Although consent increases autonomy, it places a higher burden on patients

❖An Opt-in system may introduce bias by only including a small portion of the population

❖Who would opt in? Who would be left out?

❖Underserved and seldom heard populations might be missing from the data.

For more information on the "national data opt-out" and to register your preference: https://www.nhs.uk/your-nhs-data-matters/


WATCHDOG
OVERSIGHT of Guardrails
WHO is accessing data?
WHAT data is being accessed?
HOW are outputs of data analysis being shared?
Health / Social Care
University
Approved Analysts
DATA ACCESS REQUEST
GUARDRAIL
DATA SAFE HAVEN
GUARDRAIL
Health / Social DASHBOARDS
Aggregated Anonymous
Health / Social Commissioning Service Development
RESEARCH PAPER
brighton and sussex medical school

What types of datasets are available and how long does it take/how much does it cost?

bsms.ac.uk/research/primary-care-and-population-health/primary-care/arc-data-science-hub/index.aspx

Google Scholar | My Course and Mo... | Abstract_MRC part... | Adult Autism Test R... | School Grant Applic... | 22/136 Public acce... | GitHub - Tws21/ms... | SARA: Semi-Autom...

Home | About | Undergraduate | Postgraduate | Research | Living here

ARC DATA SCIENCE HUB

- About this resource
- Datasets
- Data hub clinics
- Talking to the public about data
- Meet the team
- News, publications and events
- Papers of interest
- Contact us

DATASETS

Here you will find information about both national and regional datasets, along with resource packs containing information on access pathways, costs, training requirements, user access agreements and more. Where possible, we have included a data dictionary for the dataset so that you can explore what each dataset holds and decide whether it is an appropriate data resource for your research.

If you'd like to speak to the team for further guidance, join one of our fortnightly Data Hub Clinics!

NATIONAL DATASETS >

REGIONAL DATASETS >

Katie's resource packs:

- Key contact details
- Data dictionary (what variables dataset contains)
- How to access data – data environment and application process
- Contracts and data sharing agreements
- Costs/charges
- Training/accreditation needed.

- Please let us know if you want us to make a resource pack for another dataset!

ARC Data Hub Resource

NIHR | Applied Research Collaboration Kent, Surrey and Sussex

brighton and sussex medical school

Hospital Episode Statistics (HES)

Key contact details:

To discuss any aspect of the HES dataset you can contact the NHS Digital team via phone 0300 303 5678 or email enquiries@nhsdigital.nhs.uk

HES data dictionary:

HES is a data warehouse containing details of all admissions, outpatient appointments and A&E attendances at NHS hospitals in England. HES data covers all NHS CCGs in England, including: private patients treated in NHS hospitals, patient's resident outside of England, and care delivered by treatment centres funded by the NHS.

Each HES record contains a wide range of information about an individual patient admitted to an NHS hospital, including: clinical information about diagnoses and operations; patient information such as age group, gender and ethnicity; administrative information such as dates and methods of admission and discharge; geographic information such as where patients are treated and area where they live.

For further detailed information, take a look at the HES data dictionary.

Accessing HES data:

HES publish a number of standard analyses on the NHS Digital website. Data tables and analyses are free to access and can be downloaded and published from the data catalogue. Find monthly HES publications and the catalogue here.

Given the huge number of combinations of fields within HES, it is only possible for HES to publish a limited number of analyses. If the data you're looking for can't be found via the catalogue, you will need to apply for other standard and bespoke extracts.

HES data can be accessed through the following methods:

- The Data Access Environment – approved users can access NHS data to perform analysis. DAE hosts some of the world's leading analytics tools, for example

brighton and sussex medical school

Can we link NHS data to our research datasets?

If you want to follow up patients recruited in a study you need:

- Funding for accessing/extracting individual patients records
- Individual patient consent (use data under lawful basis of consent)
- To approach individual trusts or providers to give you the patients' data (via CRN)

Data is robustly de-identified for patient privacy and to comply with the law.

Therefore: using these datasets to follow up a recruited research cohort is not possible.

- Exception: where permission is granted to hold patient identifiers securely: e.g. NHS number, then linkage might be possible. NHS Digital is an example.
- Research cohorts (e.g. ALSPAC) now being linked to their medical records:
    - UK Longitudinal Linkage Collaboration

UK Longitudinal Linkage Collaboration

COVID-19 Longitudinal Health and Wellbeing National Core Study

HOME ABOUT US THE PUBLIC FOR RESEARCHERS FOR STUDY PARTICIPANTS PARTNER STUDIES NEWS

The UK Longitudinal Linkage Collaboration (UK LLC) has been set up to bring together information from longitudinal study volunteers with their routine records. This is being done in a secure way to help researchers work to improve health and wellbeing throughout and beyond the COVID-19 pandemic.

A new collaboration for public benefit

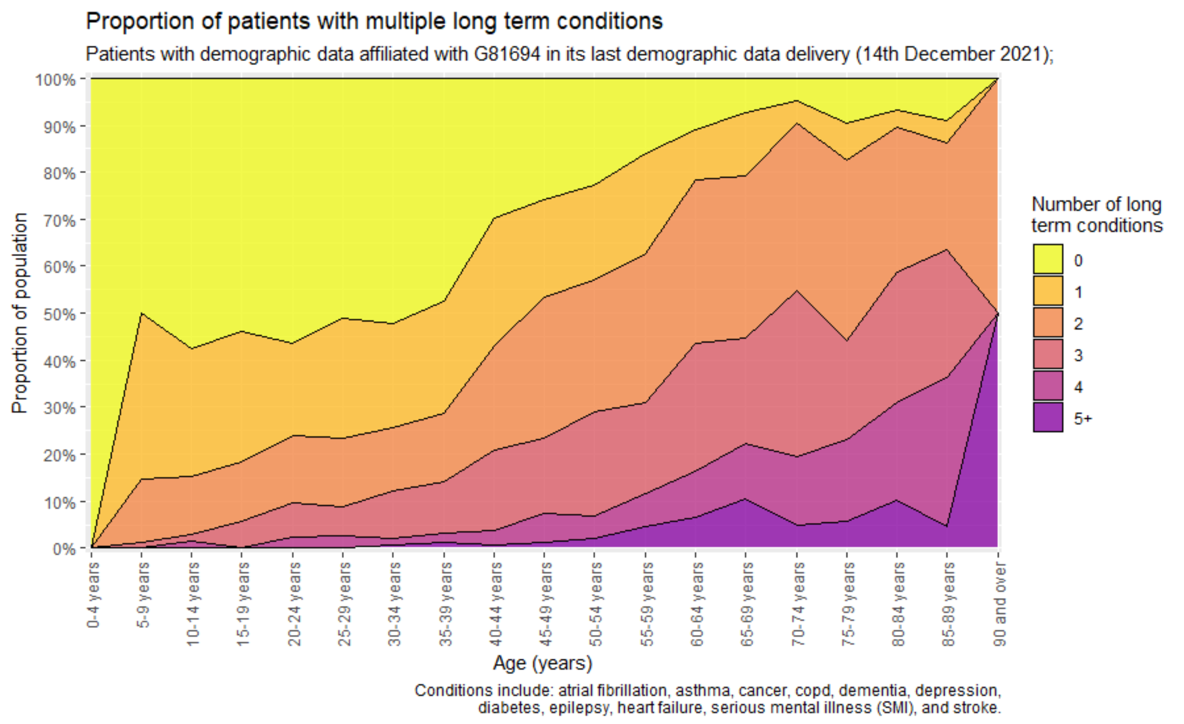
Director of the UK LLC, Andy Boyd, is giving a talk at BSMS on 3rd July at 9.30am – all welcome

[Please book your place via Eventbrite](),

brighton and sussex medical school

Example studies

Can datasets be useful for understanding inequalities?



98% of population registered with general practice

97% of Sussex general practices signed up to supply data into the Sussex Integrated Dataset

That is, good coverage of seldom heard groups.

<u>Collaboration with East Sussex County Council:</u>

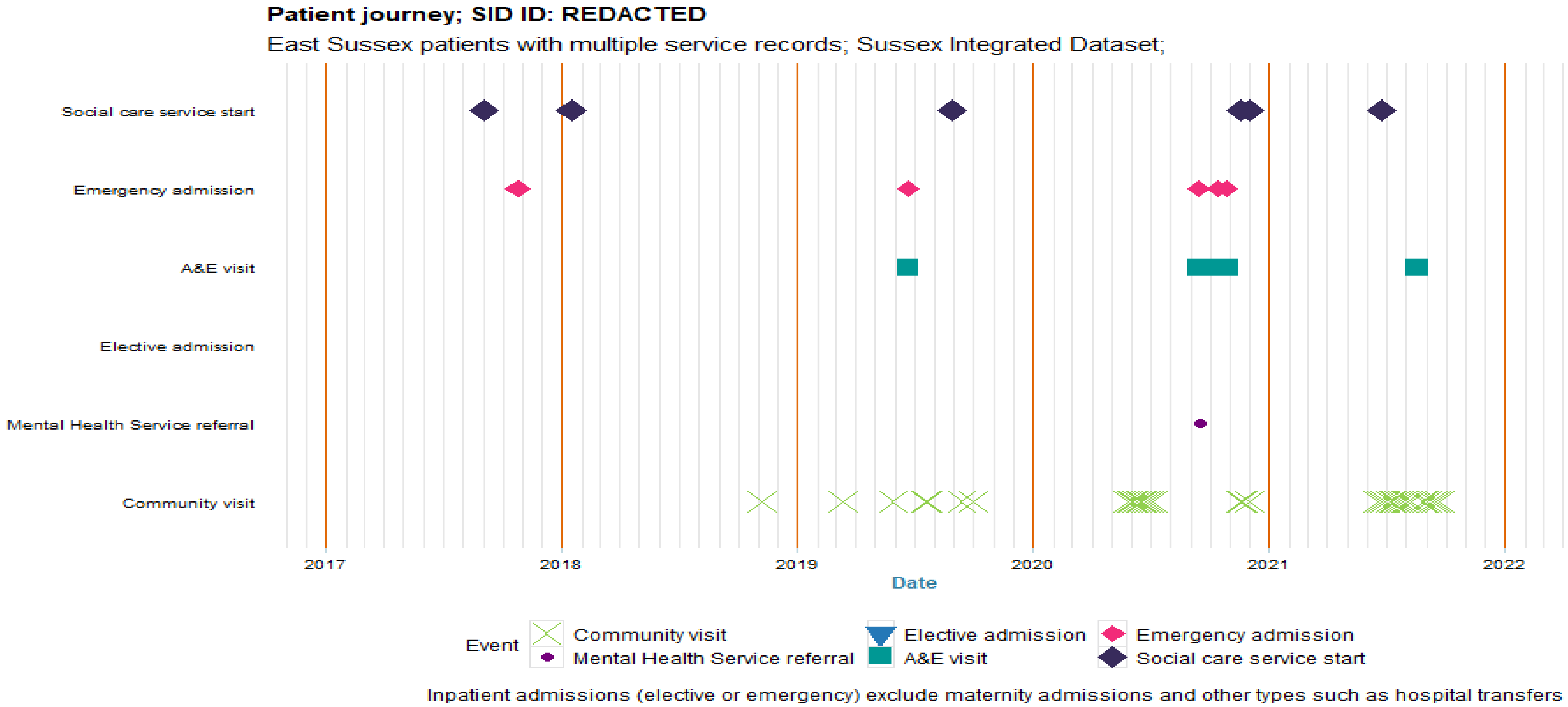
Identifying number of comorbidities by age, stratified by GP practice

Can locate areas with **higher burden of multi-morbidity at younger ages**

See full paper at https://www.mdpi.com/2078-2489/14/2/106

brighton and sussex medical school

Tracking multiple service use over time



ARC Data Science Project - Risk Stratification

The effect of multimorbidity on diagnostic interval for lung cancer and mesothelioma: a cohort study using primary care data.

- Cancer diagnosis may be delayed by the presence of comorbidities
- Two mechanisms suggested in study of colorectal cancer:
  - ***Competing demand***
    - Unrelated to the cancer but place demand on the clinician's time
  - ***Alternative explanation***
    - Conditions provide alternative explanations for cancer symptoms and delay cancer investigation
  - Both classes of comorbidities delayed diagnosis of colorectal cancer, longest delay for inflammatory bowel disease in other studies (26 days)
- We investigated effect of "competing demand" and "alternative explanation" conditions on **lung cancer/mesothelioma** diagnosis.
- Also interested in measures of inequality (IMD) and behavioural variables (smoking, BMI, alcohol drinking)

Datasets used

Clinical Practice Research Datalink

- Clinical and referral files
- Additional files
- Therapy files
- Test files
- Patient files

Linked datasets

- Hospital episodes statistics
- Cancer registry
- Index of Multiple Deprivation
  - Based postcode of GP practice
- ONS death data

Definition of key variables

Diagnostic interval – time between first presentation in primary care with a relevant symptom and date of cancer diagnosis

Date of cancer diagnosis – defined as the earliest date of a diagnostic code for lung cancer/mesothelioma in the primary care records (Read codes) or the hospital episodes statistics or cancer registry data (ICD-10 codes)

Presenting symptom – earliest date of a symptom code (within 12 months prior to diagnosis) for one of 8 relevant signs/symptoms (eg. haemoptysis, cough, dyspnoea) in the CPRD primary care data

- Code lists defined by searching code browsers, reviewing published code lists, discussion with a GP
- Most symptoms defined by Read Codes in clinical/referral files
- Thrombocytosis also defined by test result indicating raised platelets in test files

Also considered consultation frequency from 24-12m pre-diagnosis and in the 12m before diagnosis.

Comorbidities

- *Alternative explanation* conditions
  - Asthma, COPD, ACE inhibitor prescription, chronic fatigue syndrome
- *Competing demand* conditions
  - CHD, heart failure, hypertension, depression/anxiety, chronic kidney disease, osteoporosis, diabetes, epilepsy, serious mental illness
- Used presence of diagnostic Read code at any point in medical record in clinical/referral files
  - Depression and anxiety restricted to last 3 years
  - ACE inhibitor prescription from relevant product codes in therapy files in last 2 years

Behavioural variables

Smoking status

- Grouped as ever versus never smoker
- Read codes for smoking status in clinical files
- Information in additional files
- If "non-smoker" check for any previous codes for ex-smoker or current smoker

Body mass index

- Grouped as underweight, healthy weight, overweight, obese
- Read codes for BMI category in clinical files
- Data on measured height, weight, BMI in additional files
- Implausible weight and BMI values excluded

Alcohol drinking

- Grouped as current, former and never drinker
- Read codes for drinking status in clinical files
- Data on units consumed in additional files
- If "never drinker" check if in conflict with units consumed or previous codes suggesting former drinker.

Most recent data used in preference excluding 6 months prior to diagnosis

Analysis methods and results

Data were analysed by multiple linear regression with model selected by stepwise selection based on AIC.

Complete data available for 10432 lung cancer/mesothelioma cases

Diagnostic interval was longer among patients with “alternative explanation” condition

- 27.7 (95% CI 22.9, 32.4) days longer if one condition
- 72.1 (65.7, 78.4) days longer in 2 or more conditions
- Longest delay for a single condition was for COPD (60.0 days)

Number of “competing demand” conditions did not remain in model but consultation frequency was strongly positively associated with diagnostic interval

Diagnostic interval also increased by 12.7 (6.0, 19.4) days in ever- versus never-smokers and by 24.5 (14.2, 34.9) days in those in underweight patients versus those in normal weight range

How can diagnostic delays be reduced in these patients?

Lessons learned

- Allow enough time for data preparation!
- Datasets are very large, may cause problems with some software
- Check all code lists
  - Errors may exist in published code lists
- Dealing with missing data
  - May have large amounts eg. for BMI, ethnicity
  - Data unlikely to be MCAR
  - Complete case analysis? Multiple imputation?
- When defining variables need to consider
  - Time period
  - Just diagnostic codes or also medications etc.?
  - How to prioritize different sources of information at multiple timepoints and deal with conflicts
    - E.g. patient coded as non-drinker but also has value > 0 for alcohol units per week
    - Most recent data for BMI, smoking etc. or average over time, or weight gain/loss?

Resources available

Clinical code list repositories

- HDR UK Phenotype library
- LSHTM Data Compass
- ClinicalCodes.org at University of Manchester

Algorithms and software packages for dealing with health data

- Algorithms used to define variables often published on e.g Github/Zenodo
- Stata commonly used (has commands for eg icd10 codes)
- R packages eg. Aurumpipeline GitHub - HFAnalyticsLab/aurumpipeline, comorbidity, rClinicalCodes

Training courses

- CPRD training (available from MHRA Administrator Area (mhra.gov.uk))
- Safe researcher training
- Courses on working with routinely collected health data at UCL and others

What resources are available to get me started and who can I ask for help?

www.bsms.ac.uk/arcdatahub

Home | About | Undergraduate | Postgraduate | Research | Living here

ARC DATA SCIENCE HUB

- About this resource
- Datasets
- Data hub clinics
- Talking to the public about data
- Meet the team
- News, publications and events
- Papers of interest
- Contact us

Welcome to the ARC Data Science Hub!

An open access resource, identifying and exploring national and regional (Kent, Surrey and Sussex) health and social care datasets. A space where data access barriers are addressed, in the hope of encouraging improved healthcare based on the real needs of everyday people as users of health and care services.

Thank you – Questions?

Contact me:

e.m.ford@bsms.ac.uk

www.bsms.ac.uk/dr-elizabeth-ford

www.bsms.ac.uk/arcdatahub

NIHR | Applied Research Collaboration Kent, Surrey and Sussex

brighton and sussex medical school