# Statistics for Oncology

A Course for Scottish Trainees

by… The Edinburgh Cancer Informatics Research Group

https://edin.ac/oncology-statistics

# Data Management for audit and research

THE UNIVERSITY of EDINBURGH

NHS Lothian

ECi Edinburgh Cancer informatics

# Content

- Excel shaming / RWD
- Information governance
- Disclosure controls
- Data standardisation

# Why this should matters to you

Most of you will:

- Do audits
- Do retrospective studies

Data choices made *early* determine:

- Validity of results
- Easier logistics → improved workflow
- Clinical relevance

**Data collection is a *clinical skill*, not just admin.**

# ‘I’ll just put it in Excel’

- It feels easier/faster
- Familiar
- No training needed
- It’s all right here...

- Rare standardisation
- No validation process
- Silent errors
- No audit trail
- Version chaos 😱

# People’s factor

- Different people interpret variables differently
- Outcomes defined inconsistently
- Results aren’t reproducible
- Reviewer questions data integrity

**Bad data ≠ bad statistics → statistics just expose bad data.**

# Sustainability

- Datasets tend to be tied to one or 2 people, maybe even one laptop
- Not scalable
- Not re-usable by other people

# Curated data should fill gaps, not duplicate what already exists…

Cancer registry

Primary care

Cancer Waiting times
Labs

Radiotherapy (ARIA)

Finances

SACT

Genetics
Death

Inpatient

Outpatient

Digital pathology… And many more!

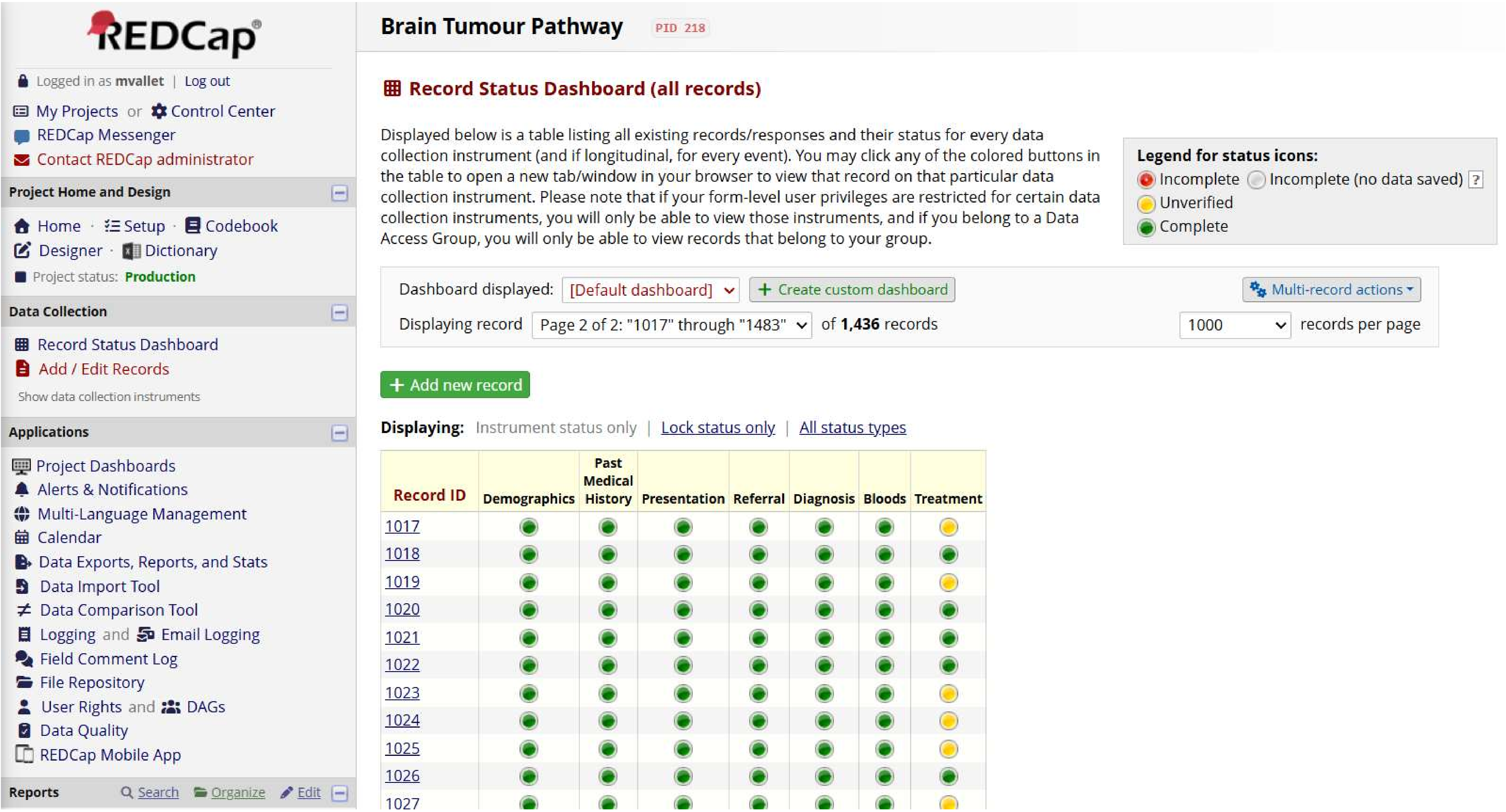
| Curated dataset | | Real world data |
|---|---|---|
| Usually collected for a project<br>Often small datasets | **Definition** | Data collected routinely as part of care<br>Examples:<br>- Cancer registry<br>- Prescribing data<br>- Inpatient data |
| Flexible<br>Can be tailored for a question | **Positives** | Data already exists<br>National standard – reproducible<br>Larger populations<br>Long follow up<br>Already QA'd<br>Usable for stat packages |
| Time consuming<br>Error prone<br>Hard to reproduce<br>Often no re-usable as is<br>Files are not usable for stat packages<br>Get lost easily | **Negatives** | May not capture all complex events<br>Requires an educational phase to get to know the data (can be quick!) |

## And if you HAVE TO curate data… Why not consider a robust system?

REDCap

Research Electronic Data Capture

- Secure, web-based data capture
- Designed for clinical research & audit
- A well known easily approved system
- Widely available in NHS / universities
- Huge online community
- Intuitive
- Free!!



Brain Tumour Pathway PID 218

Record Status Dashboard (all records)

Displayed below is a table listing all existing records/responses and their status for every data collection instrument (and if longitudinal, for every event). You may click any of the colored buttons in the table to open a new tab/window in your browser to view that record on that particular data collection instrument. Please note that if your form-level user privileges are restricted for certain data collection instruments, you will only be able to view those instruments, and if you belong to a Data Access Group, you will only be able to view records that belong to your group.

**Legend for status icons:**
🔴 Incomplete ⚪ Incomplete (no data saved) ?
🟡 Unverified
🟢 Complete

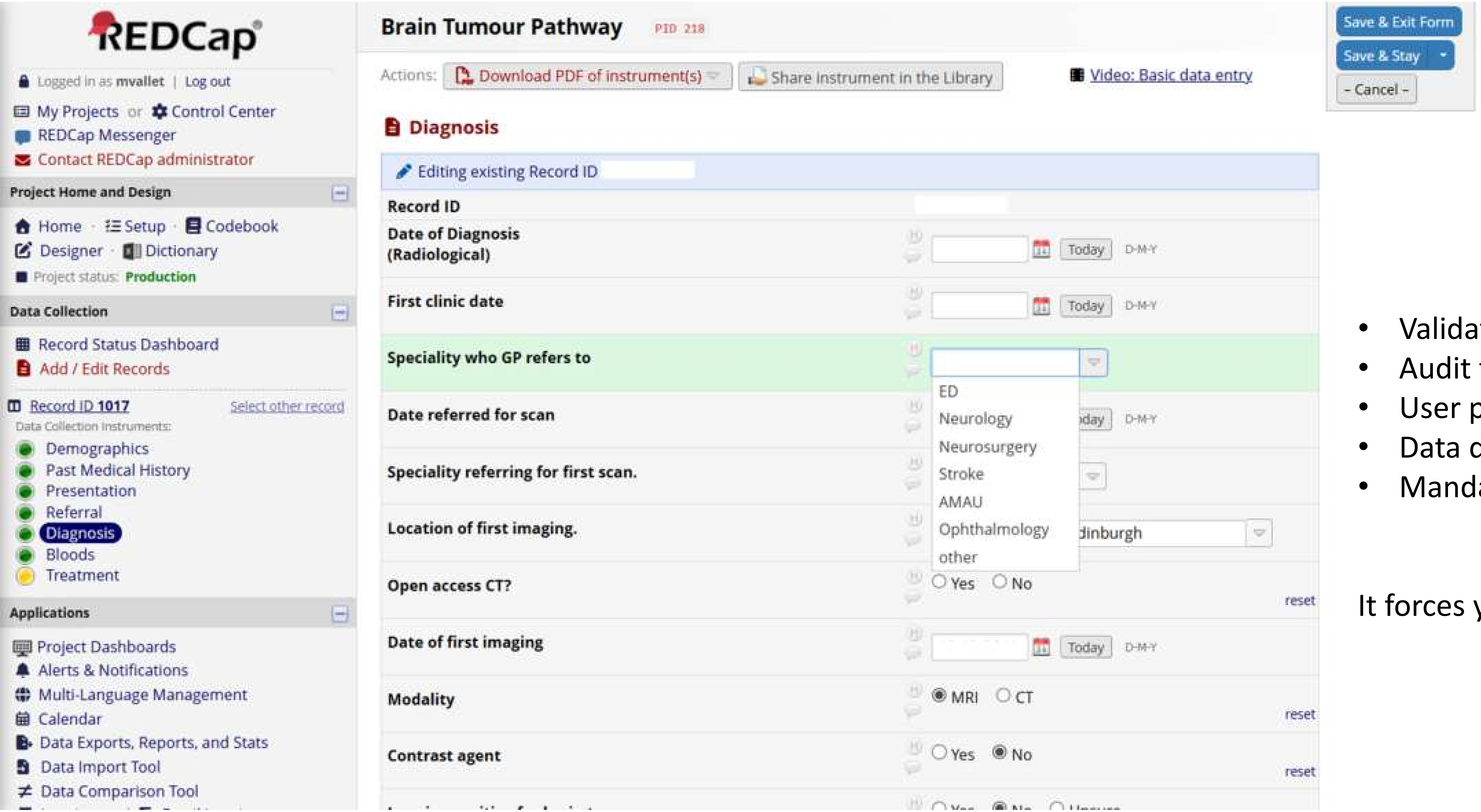
Dashboard displayed: [Default dashboard]

Displaying record Page 2 of 2: "1017" through "1483" of **1,436** records

1000 records per page

**Displaying:** Instrument status only | Lock status only | All status types

| Record ID | Demographics | Past Medical History | Presentation | Referral | Diagnosis | Bloods | Treatment |
|---|---|---|---|---|---|---|---|
| 1017 | 🟢 | 🟢 | 🟢 | 🟢 | 🟢 | 🟢 | 🟡 |
| 1018 | 🟢 | 🟢 | 🟢 | 🟢 | 🟢 | 🟢 | 🟢 |
| 1019 | 🟢 | 🟢 | 🟢 | 🟢 | 🟢 | 🟢 | 🟡 |
| 1020 | 🟢 | 🟢 | 🟢 | 🟢 | 🟢 | 🟢 | 🟢 |
| 1021 | 🟢 | 🟢 | 🟢 | 🟢 | 🟢 | 🟢 | 🟢 |
| 1022 | 🟢 | 🟢 | 🟢 | 🟢 | 🟢 | 🟢 | 🟢 |
| 1023 | 🟢 | 🟢 | 🟢 | 🟢 | 🟢 | 🟢 | 🟡 |
| 1024 | 🟢 | 🟢 | 🟢 | 🟢 | 🟢 | 🟢 | 🟡 |
| 1025 | 🟢 | 🟢 | 🟢 | 🟢 | 🟢 | 🟢 | 🟡 |
| 1026 | 🟢 | 🟢 | 🟢 | 🟢 | 🟢 | 🟢 | 🟢 |
| 1027 | 🟢 | 🟢 | 🟢 | 🟢 | 🟢 | 🟢 | 🟡 |



**Brain Tumour Pathway** PID 218

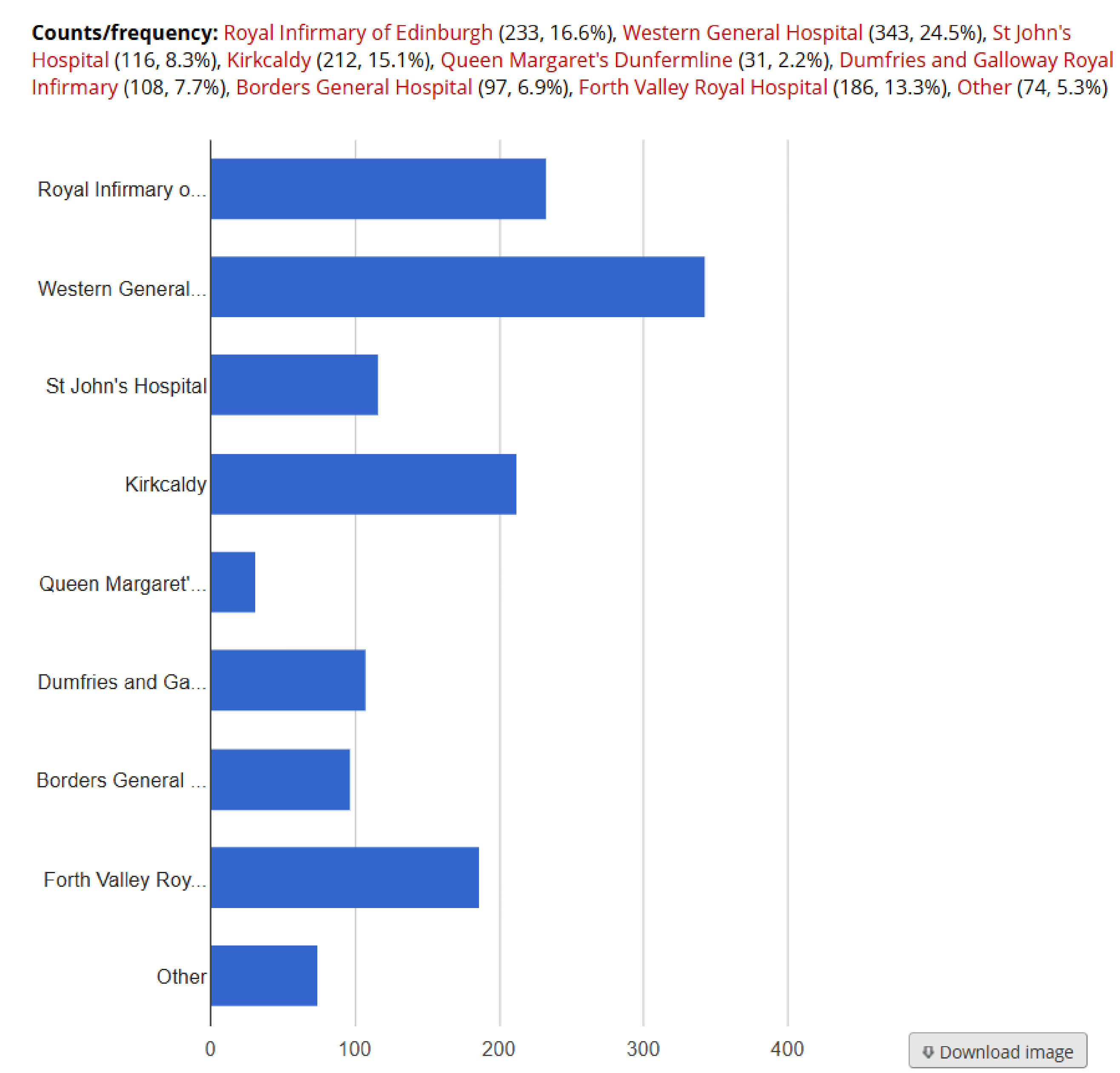
**Diagnosis**

Editing existing Record ID

| | |
|---|---|
| Record ID | |
| Date of Diagnosis (Radiological) | D-M-Y |
| First clinic date | D-M-Y |
| Speciality who GP refers to | ED<br>Neurology<br>Neurosurgery<br>Stroke<br>AMAU<br>Ophthalmology<br>other |
| Date referred for scan | D-M-Y |
| Speciality referring for first scan. | |
| Location of first imaging. | dinburgh |
| Open access CT? | ○ Yes ○ No |
| Date of first imaging | D-M-Y |
| Modality | ● MRI ○ CT |
| Contrast agent | ○ Yes ● No |
| | ○ Yes ● No ○ Unsure |

- Validation
- Audit trails
- User permission
- Data dictionary
- Mandatory fields

It forces you to be explicit!



**Counts/frequency:** Royal Infirmary of Edinburgh (233, 16.6%), Western General Hospital (343, 24.5%), St John's Hospital (116, 8.3%), Kirkcaldy (212, 15.1%), Queen Margaret's Dunfermline (31, 2.2%), Dumfries and Galloway Royal Infirmary (108, 7.7%), Borders General Hospital (97, 6.9%), Forth Valley Royal Hospital (186, 13.3%), Other (74, 5.3%)

- Quick stats/charts checks
- Easy exports

Only collect:

- Variables that aren’t in routine data (complex clinical definitions)
- Novel clinical measures
- Prospective assessments

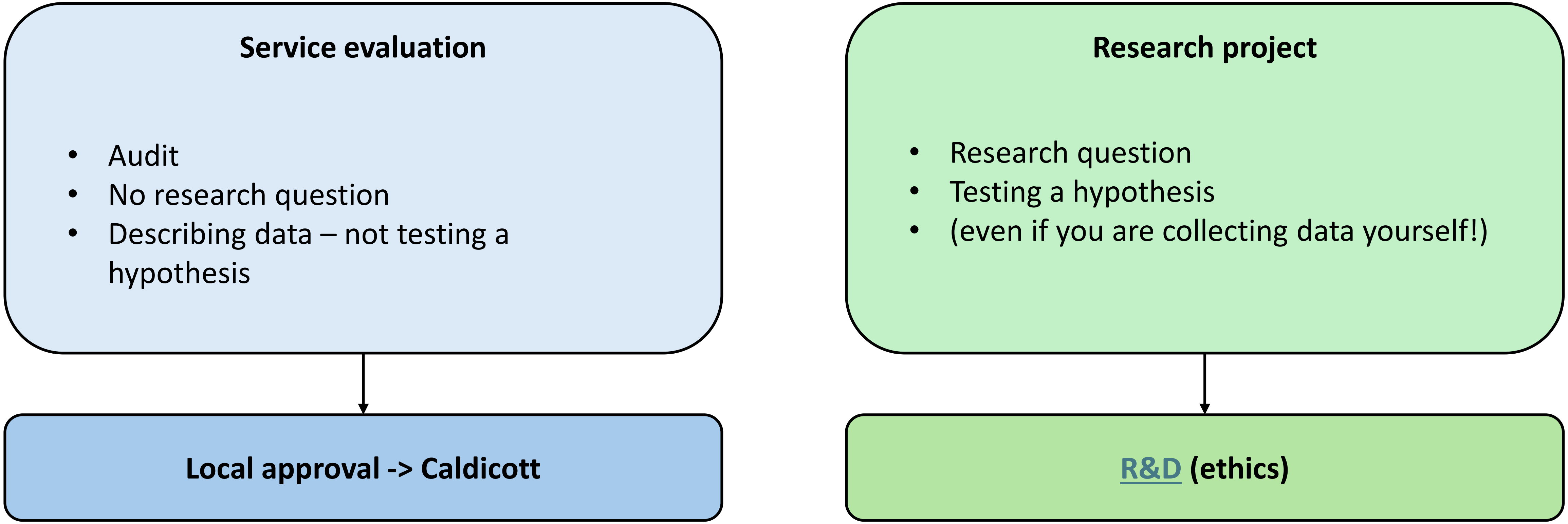
Everything else, I beg of you → RWD!

## Use of RWE for such studies is:

- Safer for the patients
- Data support if required
- Same approval pathway → tends to be quicker as provides reassurance

# Informational governance / data access

## The type of project you want to do will dictate the governance route you need to take:



**Service evaluation**

- Audit
- No research question
- Describing data – not testing a hypothesis

**Local approval -> Caldicott**

**Research project**

- Research question
- Testing a hypothesis
- (even if you are collecting data yourself!)

**R&D (ethics)**

But that can work the other way around too…

# Cancer information team – SCAN network

- You can get in touch with a team of NHS analysts who have approvals in place
- NHS Lothian, Fife, D&G, Borders data
- May save you some time and hassle…

# Disclosure controls

“Can a specific patient identify themselves in my data?”

How to protect direct identification

+

How to reduce the risk of identification

# Hide small numbers (<10)

| | N Lothian (%) n=796 | N Fife (%) n=389 | Total (%) n=1185 |
|---|---|---|---|
| Criteria 1 | 512 (64.3) | 281 (72.2) | 793 (66.9) |
| Criteria 2 | 267 (33.5) | 98 (25.2) | 365 (30.8) |
| Criteria 3 | 10 (1.3) | 8 (2.1) | 18 (1.5) |
| Criteria 4 | 7 (0.9) | 2 (0.5) | 9 (0.8) |
| Total | 796 | 389 | 1185 |

| | N Lothian (%) n=796 | N Fife (%) n=389 | Total (%) n=1185 |
|---|---|---|---|
| Criteria 1 | 512 (64.3) | 281 (72.2) | 793 (66.9) |
| Criteria 2 | 267 (33.5) | 98 (25.2) | 365 (30.8) |
| Criteria 3 | 10 (1.3) | <10 (<5) | 18 (1.5) |
| Criteria 4 | <10 (<5) | <10 (<5) | <10 (<5) |
| Total | 796 | 389 | 1185 |

Make sure the text in your paper is reflective of this + **cross check** your tables!

| | N Lothian (%) n=796 | N Fife (%) n=389 | Total (%) n=1185 |
|---|---|---|---|
| Criteria 1 | 512 (64.3) | 281 (72.2) | 793 (66.9) |
| Criteria 2 | 267 (33.5) | 98 (25.2) | 365 (30.8) |
| Criteria 3 | <15 (<5) | <10 (<5) | <20 (<5) |
| Criteria 4 | <10 (<5) | <10 (<5) | <10 (<5) |
| Total | 796 | 389 | 1185 |

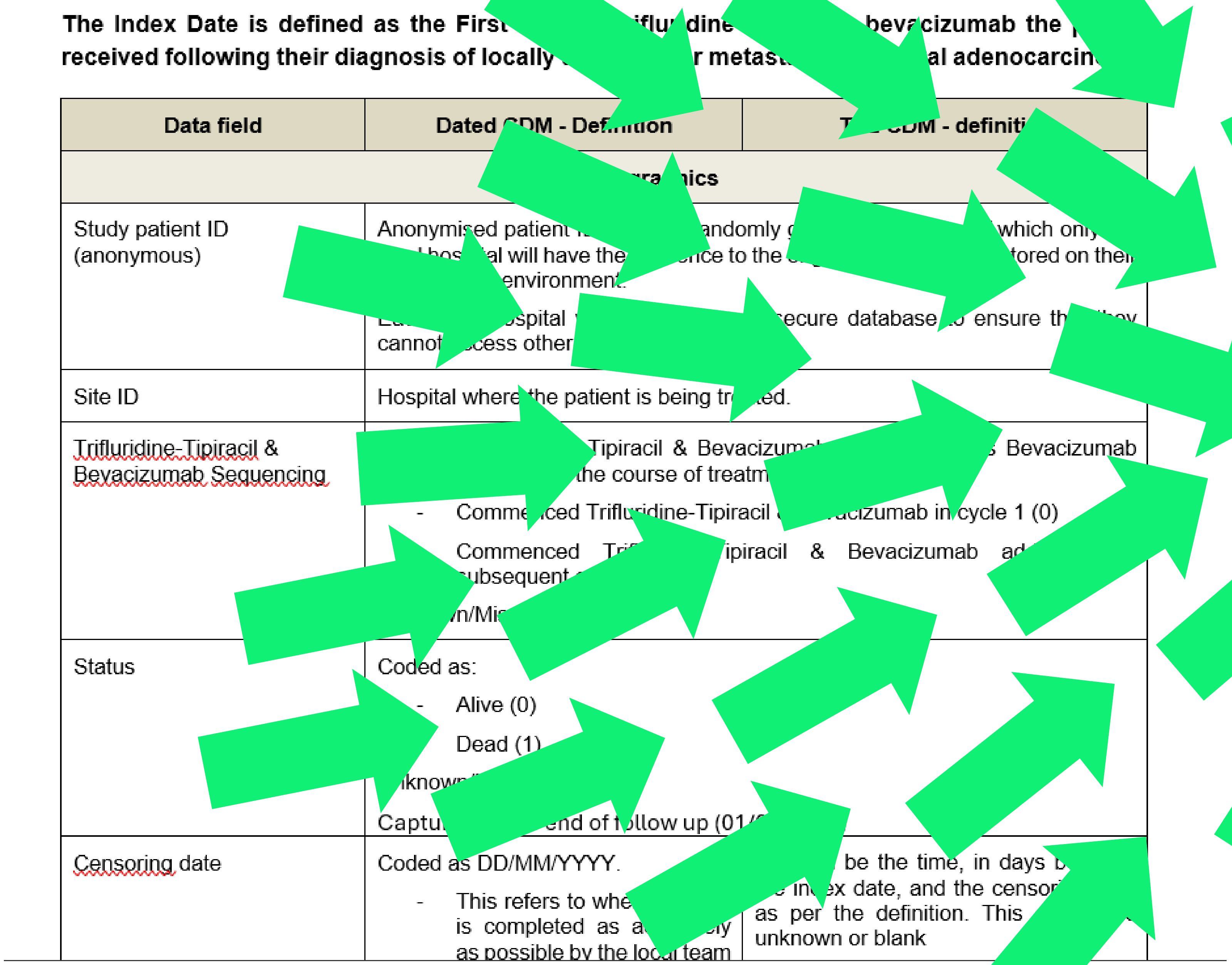
# Change data format/type

- Age ranges (cf data distribution)
- BMI/BSA ranges instead of height and weight
- Check data availability/accuracy before reporting (ethnicity?)
- Replace dates by time to events

Ask a (work) friend for a fresh pair of eyes…

# Data standardisation

# Common data mo ls



**The Index Date is defined as the First iflu dine bevacizumab the received following their diagnosis of locally r metast al adenocarcin**

| Data field | Dated CDM - Definition | T CDM - definiti |
|---|---|---|
| ra hics | | |
| Study patient ID (anonymous) | Anonymised patient andomly g which only hospital will have the ence to the o tored on their environment.<br>Ea spital ecure database to ensure th they cannot ccess other | |
| Site ID | Hospital where the patient is being tr ted. | |
| Trifluridine-Tipiracil & Bevacizumab Sequencing | Tipiracil & Bevacizuma s Bevacizumab the course of treatm<br>- Comme ced Trifluridine-Tipiracil acizumab in cycle 1 (0)<br>- Commenced Tri ipiracil & Bevacizumab ad subsequent<br>n/Mis | |
| Status | Coded as:<br>- Alive (0)<br>- Dead (1)<br>known/<br>Captu end of follow up (01/ | |
| Censoring date | Coded as DD/MM/YYYY.<br>- This refers to whe is completed as a ly as possible by the local team | be the time, in days b e index date, and the censori as per the definition. This unknown or blank |

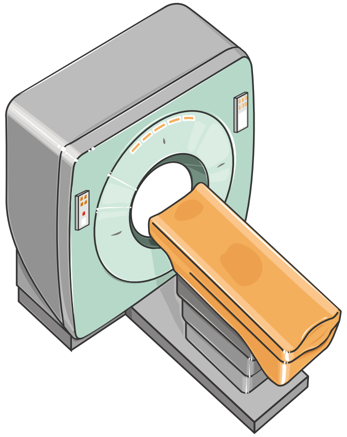
Data form
ists of c ies
Always something fo
unknown/missing (no
empty!)

- PROPER DEFINITIONS

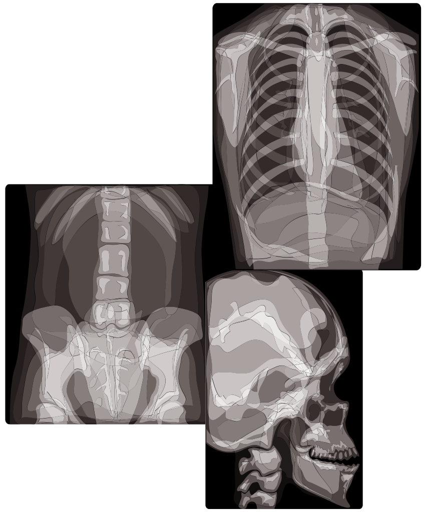
02/02/04
Pt phone GP

09/02/04
GP appointment

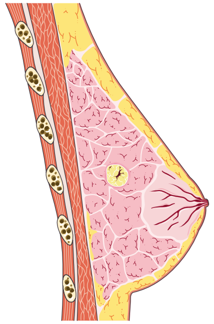
21/02/04
Sent to
hospital



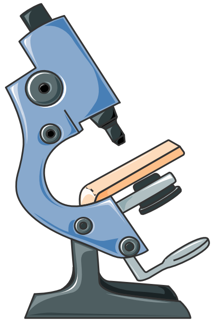
01/03/04
Early
investigations



14/03/04
Scan report
cTNM

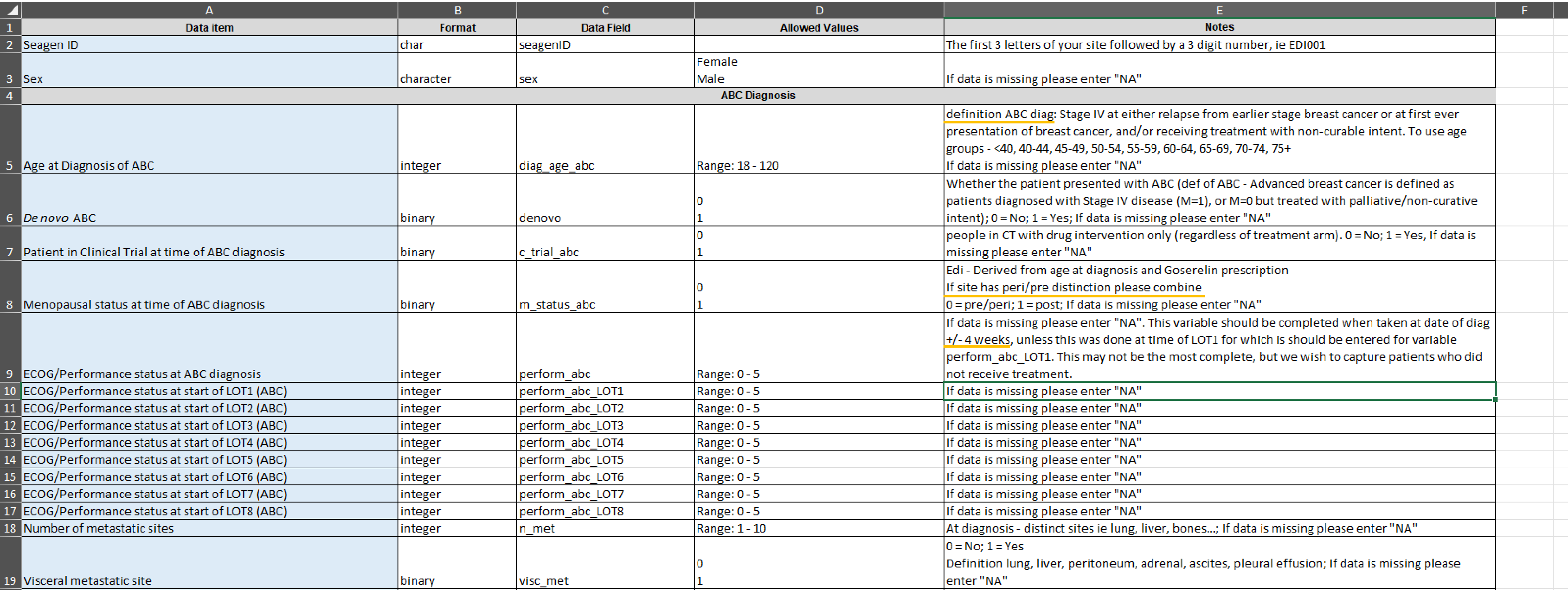


23/03/04
Surgery



30/03/04
Path report
pTNM

What is the date of diagnosis?



|  | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | **Data item** | **Format** | **Data Field** | **Allowed Values** | **Notes** |  |
| 2 | Seagen ID | char | seagenID |  | The first 3 letters of your site followed by a 3 digit number, ie EDI001 |  |
| 3 | Sex | character | sex | Female<br>Male | If data is missing please enter "NA" |  |
| 4 | **ABC Diagnosis** |  |  |  |  |  |
| 5 | Age at Diagnosis of ABC | integer | diag_age_abc | Range: 18 - 120 | definition ABC diag: Stage IV at either relapse from earlier stage breast cancer or at first ever presentation of breast cancer, and/or receiving treatment with non-curable intent. To use age groups - <40, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75+<br>If data is missing please enter "NA" |  |
| 6 | *De novo* ABC | binary | denovo | 0<br>1 | Whether the patient presented with ABC (def of ABC - Advanced breast cancer is defined as patients diagnosed with Stage IV disease (M=1), or M=0 but treated with palliative/non-curative intent); 0 = No; 1 = Yes; If data is missing please enter "NA" |  |
| 7 | Patient in Clinical Trial at time of ABC diagnosis | binary | c_trial_abc | 0<br>1 | people in CT with drug intervention only (regardless of treatment arm). 0 = No; 1 = Yes, If data is missing please enter "NA" |  |
| 8 | Menopausal status at time of ABC diagnosis | binary | m_status_abc | 0<br>1 | Edi - Derived from age at diagnosis and Goserelin prescription<br>If site has peri/pre distinction please combine<br>0 = pre/peri; 1 = post; If data is missing please enter "NA" |  |
| 9 | ECOG/Performance status at ABC diagnosis | integer | perform_abc | Range: 0 - 5 | If data is missing please enter "NA". This variable should be completed when taken at date of diag +/- 4 weeks, unless this was done at time of LOT1 for which is should be entered for variable perform_abc_LOT1. This may not be the most complete, but we wish to capture patients who did not receive treatment. |  |
| 10 | ECOG/Performance status at start of LOT1 (ABC) | integer | perform_abc_LOT1 | Range: 0 - 5 | If data is missing please enter "NA" |  |
| 11 | ECOG/Performance status at start of LOT2 (ABC) | integer | perform_abc_LOT2 | Range: 0 - 5 | If data is missing please enter "NA" |  |
| 12 | ECOG/Performance status at start of LOT3 (ABC) | integer | perform_abc_LOT3 | Range: 0 - 5 | If data is missing please enter "NA" |  |
| 13 | ECOG/Performance status at start of LOT4 (ABC) | integer | perform_abc_LOT4 | Range: 0 - 5 | If data is missing please enter "NA" |  |
| 14 | ECOG/Performance status at start of LOT5 (ABC) | integer | perform_abc_LOT5 | Range: 0 - 5 | If data is missing please enter "NA" |  |
| 15 | ECOG/Performance status at start of LOT6 (ABC) | integer | perform_abc_LOT6 | Range: 0 - 5 | If data is missing please enter "NA" |  |
| 16 | ECOG/Performance status at start of LOT7 (ABC) | integer | perform_abc_LOT7 | Range: 0 - 5 | If data is missing please enter "NA" |  |
| 17 | ECOG/Performance status at start of LOT8 (ABC) | integer | perform_abc_LOT8 | Range: 0 - 5 | If data is missing please enter "NA" |  |
| 18 | Number of metastatic sites | integer | n_met | Range: 1 - 10 | At diagnosis - distinct sites ie lung, liver, bones...; If data is missing please enter "NA" |  |
| 19 | Visceral metastatic site | binary | visc_met | 0<br>1 | 0 = No; 1 = Yes<br>Definition lung, liver, peritoneum, adrenal, ascites, pleural effusion; If data is missing please enter "NA" |  |

(Did I mention proper definitions?)

# Common data models

- OHDSI - Observational Medical Outcomes Partnership
- MEDOC - Minimal Essential Description of Cancer

# Conclusion

- Standardisation = speed
- Only curate what you need to
- Develop relationships with your IT/local ethics teams
- Improves collaboration
- Think long-term

- But be cautious of linkage, and assumptions. Analysts need clinical support, and vice versa with domain knowledge  this is never a solo job!
- Unsolicited advice (sorry)

Thank you!

maheva.vallet@ed.ac.uk
Edinburgh cancer informatics