Semi-automated conversion of several genetic test reports into a unique standard format

Dr. Francesco Santaniello, PhD

CanGene-CanVar Staff Meeting and Management Committee Meeting
10th December 2020

CanGene CanVar

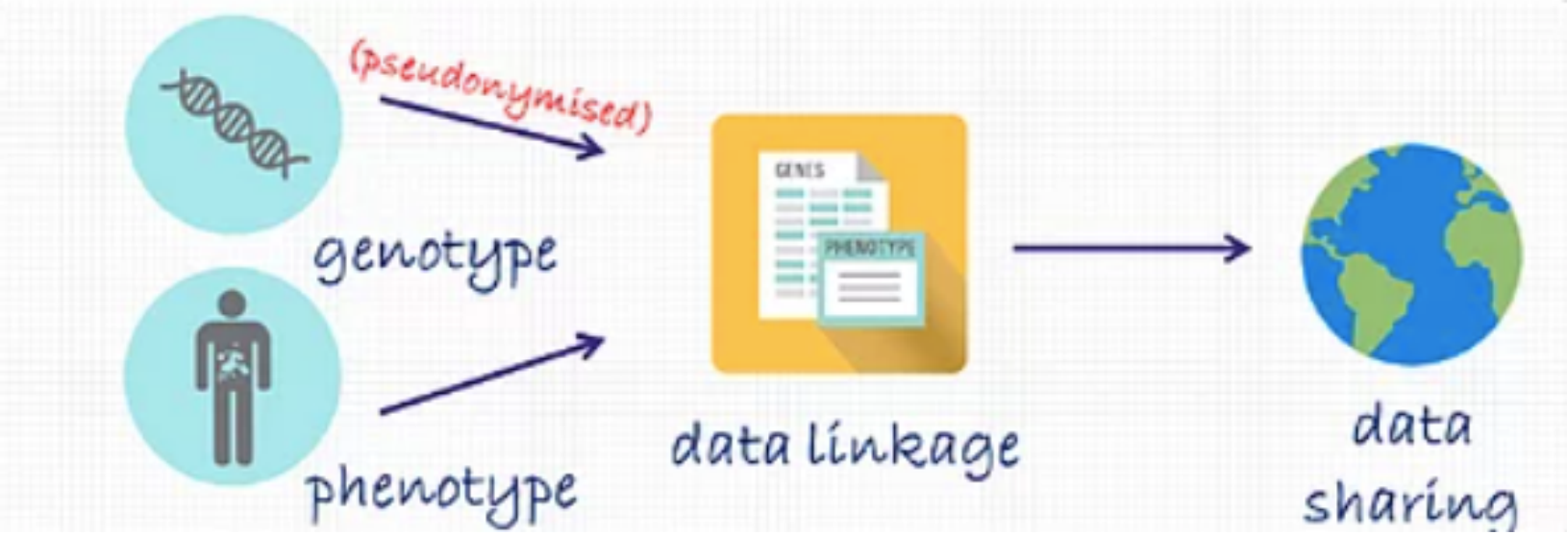
HDI healthdatainsight

Public Health England

Bowel Cancer UK

CanGene-CanVar WP1

data collection and linkage



CanGene CanVar

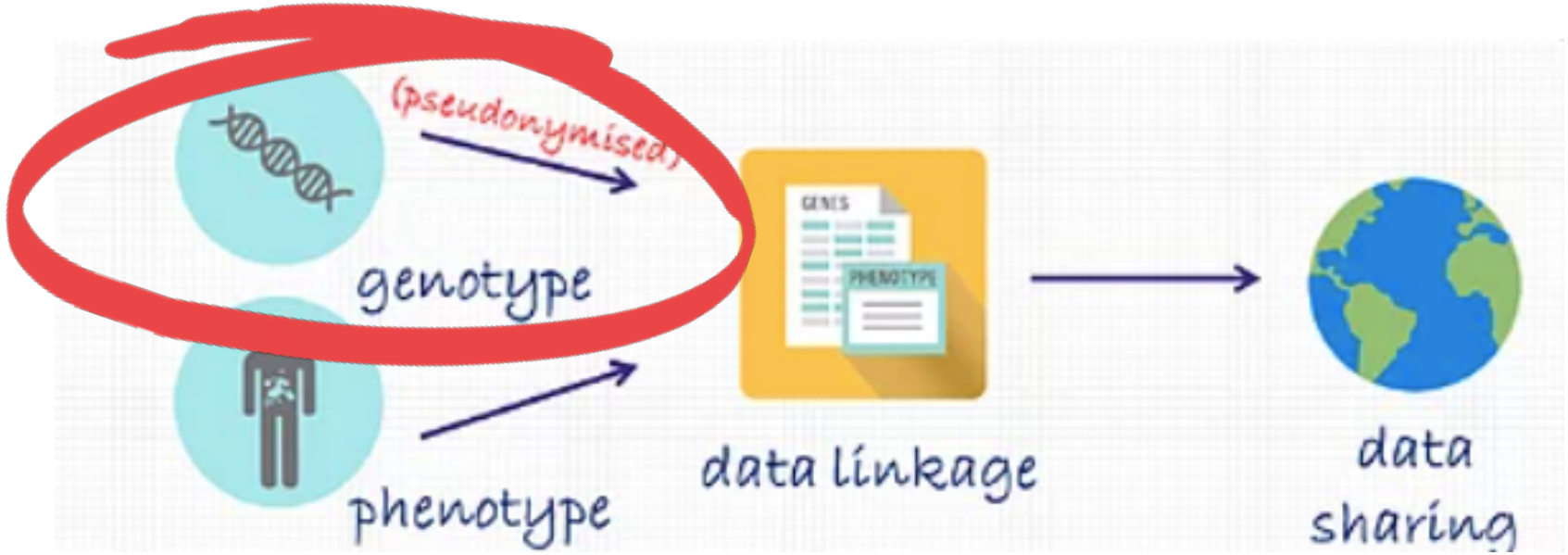
HDI healthdatainsight

Public Health England

Bowel Cancer UK

CanGene-CanVar WP1

data collection and linkage



CanGene CanVar

HDI healthdatainsight

Public Health England

Bowel Cancer UK

DATA COLLECTION

- Genetic tests from several providers are “sent to Cambridge” (uploaded on API Encore portal)

CanGene CanVar

HDI healthdatainsight

Public Health England

Bowel Cancer UK

DATA PSEUDONYMISATION

- Genetic tests from several providers are “sent to Cambridge” (uploaded on API Encore portal)
- Upon upload, patient sensitive information is pseudonymised

First Name: Jane
Last name: Smith
Address: Sesame St. 241, Cambridge
NHS Number: 123-456-789
Age: 94
Sex: F
Date of Birth: 21/04/1927
Postcode: SW1A 1AA

First Name: 38nfyepqwahdofmq39urowf
Last name: hfeidgflh298oy3nr5oe329ur
Address: 0932nyepqwhf9yh9328href
NHS Number: 832ynfeihfiduhihakjhkdgidg
Age: 9r84nyrgeihiudhhasilg84w8
Sex: hfkjfwldg983yrggwgliU84Y2
Date of Birth: mrhuf392gigeiw94hteffn938
Postcode: pnr8qhefiudhsih39984yt2095

CanGene CanVar

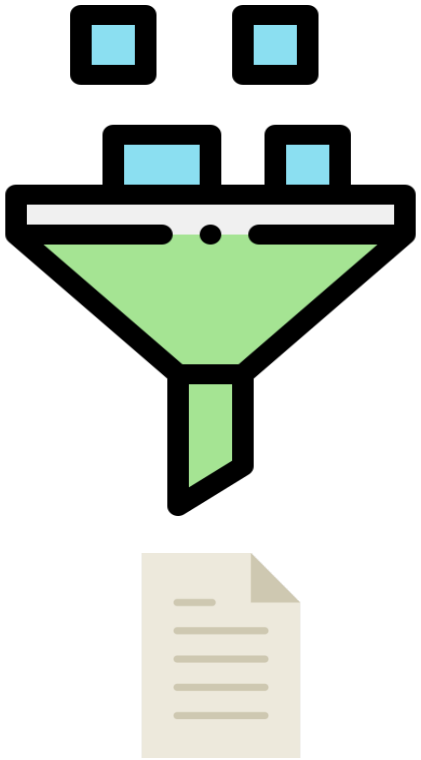
HDI healthdatainsight

Public Health England

Bowel Cancer UK

DATA STANDARDISATION – first pass

- Genetic tests from several providers are “sent to Cambridge” (uploaded on API Encore portal)
- Upon upload, patient sensitive information is pseudonymised
- Simultaneously, each report is standardised by a first pass of ad-hoc scripts, in order to map as many columns as possible by the mean of ‘*simple*’ standard yaml mappings.



```yaml
- column: hosp_no
  rawtext_name: hospitalnumber
  mappings:
  - field: hospitalnumber
```

```yaml
- column: dob
  rawtext_name: dateofbirth
  mappings:
  - field: dateofbirth
    format: %d/%m/%Y
```

```yaml
- column: prov_code
  rawtext_name: providercode
  mappings:
  - field: providercode
```

CanGene CanVar

HDI healthdatainsight

Public Health England

Bowel Cancer UK

DATA STANDARDISATION – second pass

After first pass, each record must parsed according to:

1. Their own provider rules
2. Mapped and unmapped columns
3. Presence of free-text columns
4. Set of genes being tested

CanGene CanVar

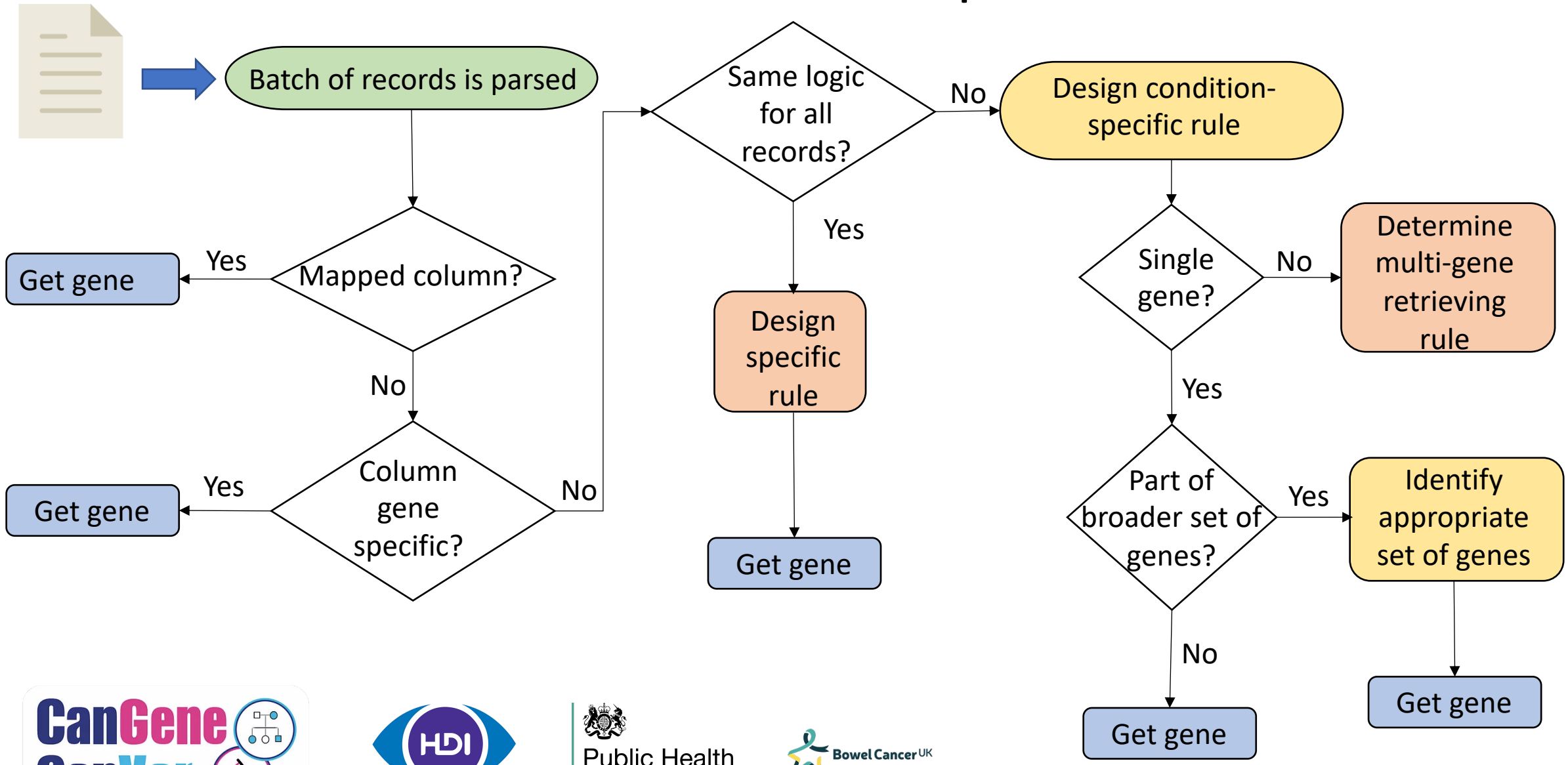
HDI healthdatainsight

Public Health England

Bowel Cancer UK

DATA STANDARDISATION – second pass

Gene extraction rules – a simplified workflow



CanGene CanVar

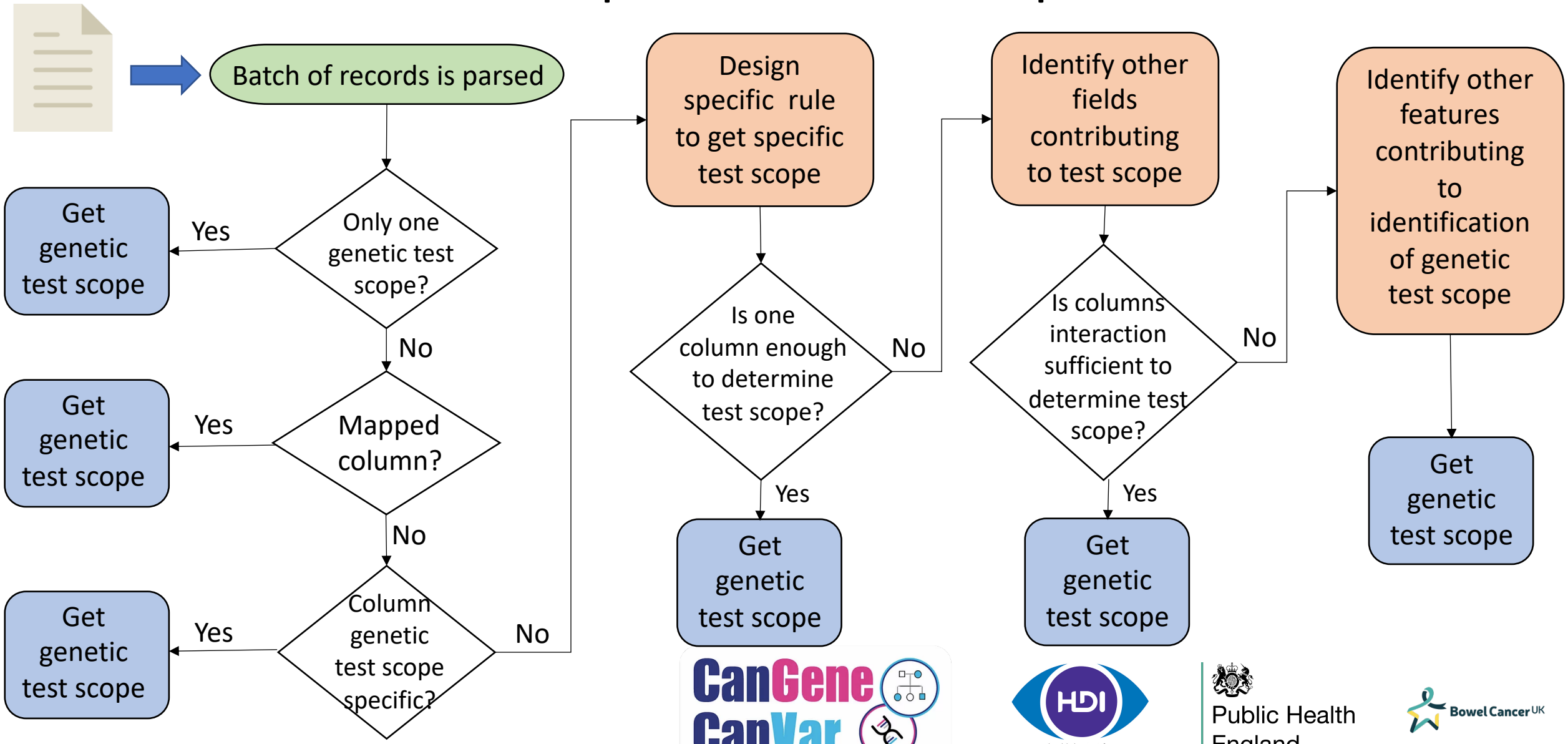
HDI healthdatainsight

Public Health England

Bowel Cancer UK

DATA STANDARDISATION – second pass

Genetic test scope extraction rules – a simplified workflow



CanGene CanVar

HDI healthdatainsight

Public Health England

Bowel Cancer UK

DATA STANDARDISATION – second pass

Real life examples

Simplest case:

- **All records are relative to full screen tests**
- **All columns are mapped, one gene per row**
- **A single rule for each column**
  - **cdna change and protein change have the same rule**

mapped:gene	mapped:genomicchange	mapped:codingdnasequencechange	mapped:proteinimpact
8	Chr13.hg19:g.32915160	c.[6668T>C]+[=]	p.[Phe2223Ser]+[=]
8	Chr13.hg19:g.32903578	c.[632-2A>G]+[=]	NA
8	Chr13.hg19:g.32913774	c.[5282G>A]+[=]	p.[Gly1761Glu]+[=]
8	Chr13.hg19:g.32913794	c.[5303_5304delTT]+[=]	p.[Leu1768Argfs*5]+[=]
7	Chr17.hg19:g.41251834	c.[505C>T]+[=]	p.[Gln169*]+[=]
8	Chr13.hg19:g.32910676	c.[2186_2190delTAAAA]+[=]	p.[Ile729Argfs*20]+[=]
8	Chr13.hg19:g.32911181	c.[2689G>C]+[=]	p.[Glu897Gln]+[=]

DATA STANDARDISATION – second pass

Real life examples

Intermediate case:

- **Genetic test scope to be identified by two columns**
- **Gene, cdna change and protein impact declared in a single column**
  - **Protein impact and cdna change have different formats throughout the batch**
  - **Genes not always declared – need to be extrapolated from another column**
  - **Abnormal and normal tests mixed up together**

raw:genotype	raw:genetictestscope	raw:karyotypingmethod
SMAD4:c.[1573A>G];[=] p.[(Ile525Val)];[(=)] MUTYH: c.[1014G>C];[=] p.[(Glu338His)];[(=)] -See below	Colorectal cancer panel	Full panel
No pathogenic mutation detected	Colorectal cancer panel	Full panel
No pathogenic mutation detected	Colorectal cancer panel	Full panel
MSH6 c.[2194C>T];[=] p.[(Arg732*)];[(=)]	Colorectal cancer panel	MLH1 MSH2 & MSH6
No pathogenic mutation detected	Colorectal cancer panel	APC & MUTYH
MUTYH: c.536A>G(;)1187G>A, p.(Tyr179Cys)(;)(?)	R209 :: Inherited colorectal cancer (with or without polyposis)	R209.1 :: NGS - APC and MUTYH only
MSH6 c.[1382T>C];[=], p.[(Phe461Ser)];[(=)]	R210 :: Inherited MMR deficiency (Lynch syndrome)	R210.2 :: Unknown mutation(s) by Small panel

CanGene CanVar

HDI healthdatainsight

Public Health England

Bowel Cancer UK

DATA STANDARDISATION – second pass

Challenging case:

- **Genetic test scope to be identified a column. Several labels for genetic test scopes**
- **Gene, cdna change and protein impact declared in a single column free-text column**
  - **Different formats for cdna change and protein impact**
  - **Multiple variants per gene in abnormal tests**
  - **Multiple genes in abnormal tests (not shown)**
  - **Abnormal and normal tests mixed up together**
  - **'Baits' in free text – e.g. variant identified as absent, or genes and variants being quoted from literature but not actually tested**

raw:genetictestscope	raw:report
Confirmation	Sequence analysis confirms that this patient is heterozygous for the familial pathogenic MSH2 mutation c.1609A>T (p.Lys537X). This result is consistent with this patient's affected status. Testing for this mutation is now available to this patient,Äôs relatives as appropriate. Please note that analysis of MSH2 exon 10 was previously done using SSCP and the familial mutation was not detected (see report dated 06/09/05). This is likely due to the reduced sensitivity of SSCP compared with sequencing.
Diagnostic	This patient has been screened for mutations in all coding exons of MLH1, MSH2 and MSH6 by sequence analysis [see notes below]. No pathogenic mutations were identified. MLPA analysis of MLH1, MSH2 and MSH6 showed no evidence of a deletion or duplication within these genes.
Predictive	Analysis indicates that the familial MSH2 sequence variant c.2288C>T (p.Ala763Val) is absent in this patient. Assuming that this variant represents the pathogenic change within this family, this result significantly reduces her risk of developing MSH2-associated cancers. This result does not affect her risk of developing other familial or sporadic cancers.
Diagnostic	Analysis indicates that this patient is heterozygous for the sequence variants c.1387-8G>T and c.1662-9G>A in MSH2. Both of these changes are listed as unknown variants on the LOVD database* and splice site prediction software** used in this laboratory did not suggest that these variants would have a deleterious effect. Evaluation of the available evidence suggests that these variants are likely to be benign. Please note that we did not confirm the presence of the c.1662-9G>A variant by Sanger sequencing as there was not enough DNA to carry out analysis.
Diagnostic	Analysis indicates that this patient is heterozygous for the sequence variant c.2259delT (p.Phe753fs) in exon 19 of MLH1.This frameshift mutation occurs near the end of the MLH1 gene and therefore may not lead to nonsense-mediated decay. However, if nonsense-mediated decay did not occur this variant would cause alteration of the last four amino acids of the MLH1 protein. These last four amino acids show 100% conservation across species and there is significant evidence in the literature that residues 492-756 are involved in the binding of MLH1 to PMS2. Evaluation of the available evidence therefore indicates that this variant is highly likely to be pathogenic. This result is consistent with the patient's affected status, and the patient is at high risk of developing further HNPCC-related cancers. This result may have important implications for other family members and testing is available if appropriate. We recommend that those relatives are referred to their local Clinical Genetics department. *Please note: no result was obtained for MLPA P003 (MLH1 and MSH2). Please inform us if testing for this assay is still required.
Diagnostic	This patient has been screened for MLH1, MSH2 and MSH6 mutations by sequence analysis and MLPA. This patient is heterozygous for the MSH6 sequence variant c.3024C>T (p.=). Evaluation of the available evidence suggests that this variant is likely to be benign as it is not predicted to affect splicing of MSH6.

DATA STANDARDISATION – results

Each record is converted in a standard output - ready to be linked to the registry

pseudo_id1	pseudo_id2	Codingdna sequencechange	gene	proteinimpact	provider	moleculartestingtype	genetictestscope
ur823nrioewu	t8937nefw;o9238e2	c.9433G>C	8	p.Val3145Leu	RQ3	2	Full screen BRCA1 and BRCA2
p92n83fdhasuf	jdhfgdis86y34yf823yw	c.7141C>T	8	p.Pro2381Ser	RQ3	1	Full screen BRCA1 and BRCA2
h8142g8e233d	wnsugf87gslsruyghsks	c.7679_7680del	8	p.Phe2560SerfsTer5	RQ3	2	Targeted BRCA mutation test
15f7af25fc2c6	mhfh927grisjsj337400	c.4065_4068del	7	p.Asn1355LysfsTer10	RQ3	1	Full screen BRCA1 and BRCA2

CanGene CanVar

HDI healthdatainsight

Public Health England

Bowel Cancer UK

DATA STANDARDISATION – results

An example of what we can extrapolate from a standardized format

A clean summary of variant counts / variant frequencies in full screen tests

dna	impact	gene	variantclass	rq3	rvj	rgt	rr8	rtd	rx1	rnz	rcu
c.4065_4068del	p.Asn1355Lysfs	BRCA1	4,5	11	2	7	35	9	23	9	6
c.6275_6276del	p.Leu2092ProfsTer7	BRCA2	3,5	17	0	3	8	4	9	23	3
c.3756_3759del	p.Leu1252fs	BRCA1	5	22	0	5	4	4	5	15	1
c.68_69del	p.Glu23ValfsTer17	BRCA1	2000,5	7	0	1	11	3	14	6	7

CanGene CanVar

HDI healthdatainsight

Public Health England

Bowel Cancer UK

ACKNOWLEDGEMENTS

Dr. Steven Hardy, Dr. Fiona McRonald, Oliver Tulloch, Dr. Brian Shand, Shilpi Goel

Dr. Joanna Pethick, Dr. Eleni Sofianopoulou

PHE NDR IT Team

Health Data Insight Staff

Public Health England Staff

Dr. Jem Rashbass

Professor Sir John Burn

Bowel Cancer UK
Beating bowel cancer together

…and you all for the attention!

CanGene CanVar

HDI healthdatainsight

Public Health England

Bowel Cancer UK