Supplemental Methods for manuscript entitled, “Integrated epidemiogical and genomic data yields insights into the relationship between precancer and cancer states of the oesophagus”

Authors: S. A. Zamani, L. Wu, E. L. Black, A. Bartram, A. W. T. Ng, M. Secrier, D. Jacobson, G. Devonshire, N. Grehan, B. Nutzinger, A. Freeman, A. Miremadi, M. O'Donovan, A. M. Frankell, S. Killcoyne, OCCAMS Consortium, H. G. Coleman, and R. C. Fitzgerald.

Contents

Esophageal adenocarcinoma cohort

Selection of cases

The inclusion criteria for the OCCAMS study select patients with a confirmed diagnosis of adenocarcinoma of the esophagus, stomach, and gastro-esophageal junction who were fit enough for treatment on a curative pathway, which was generally neo-adjuvant chemotherapy and surgery. For these cases, we aimed to collect pre-treatment samples for sequencing but where this was not possible a sample was taken from the surgical resection specimen. For patients with early disease, treatment comprised endoscopic therapy (endoscopic mucosal resection with or without radiofrequency ablation). A small number of advanced-stage patients were included who were initially deemed to be curative but in whom the full staging showed more advanced disease not suitable for a curative pathway.

Comprehensive clinical research guidelines were developed in the Fitzgerald Research Group and followed by trained clinical and research staff at all OCCAMS study centers. At each study center, eligible esophageal adenocarcinoma (EAC) patients were identified and approached regarding participation in the OCCAMS study and their desire to join as participants. Alternatively, an OCCAMS research staff in the clinic asked the patient's permission to be contacted with more information via mail and a follow-up phone call. Patients had the opportunity to ask questions and think about their involvement by talking to their GP, for example, and could return the signed form later. Consent was obtained from patients to contact their GP to inform them of their participation in OCCAMS and to obtain relevant medical information from the cancer registries and other NHS data controllers.

Patients with pathologically assessed tumors and diagnosed with adenocarcinoma of the esophagus between 2002-2022 were included. Patients with tumor histology other than EAC

(were excluded, and a majority (n=233, 67%) were ESCC cases. Furthermore. Patients with 'open & shut' surgery with more advanced disease than expected were also excluded. This was because a tumor sample was generally not collected for these patients, therefore tumor phenotype ascertainment would have not been possible. In addition, little data was recorded on the baseline questionnaire forms for such patients. A small number of cases (n=22, <1%) were missing age or sex and these were excluded.

Pathology review

A strict expert pathology review was performed for all cases. At least two pathologists reviewed each EAC case: one pathologist from the referring study center and another pathologist from the OCCAMS central study center at Cambridge University Hospitals who had more than 20 years of experience in upper GI cancer. Tumors were staged based on the UICC/AJCC tumor, Nodes and Metastases (TNM) Guidelines (7th edition)[1] . The T, N and M stages were assigned using the available information in the patient's medical records including clinical chart notes, endoscopic ultrasound, positron emission tomography, endoscopic mucosal resection and histopathological reports following surgical resection. We used the most advanced stage prior to or at the time of surgery for patients who received neoadjuvant therapy.

The presence of BE adjacent to EAC for OCCAMS cases based on endoscopic (macroscopic) visual changes observed at pre-staging evaluation with pathology review showing IM at the time of surgical resection which was assessed by expert GI pathologists of the recruiting OCCAMS sites. IM was also identified in cases without macroscopic evidence of BE upon expert review of the pathology specimen. All pathologists followed a

specific synoptic report proposed by the College of American Pathologists. Additionally, pathologists followed the OCCAMS study protocol which required thorough assessment for BE in the proximal and distal resection margins and tumor. Tumor sampling was done for all borders of the resected tumor and the tumor bed to minimize sampling error. The number of biopsy specimens varied based on tumor size.

Clinical data collection

Trained research staff collected baseline characteristics using chart review or during structured face-to-face interviews using a uniform case report form (CRF) across the 25 sites in the OCCAMS Consortium. All covariates used in the analysis originated from the study CRF. Patient baseline characteristics were collected on demographic, anthropometric, and environmental exposures. Weight and height were measured objectively at the baseline visit or from the next closest record to the baseline. Overall survival time (in years) was calculated from the date of diagnosis to the date of death or the date the patient was last seen in the clinic. Vital status was ascertained from all-cause mortality. All patients who consented to participate provided the minimum reporting standard which required demographic and clinical details.

Research staff transcribed and entered data captured on the OCCAMS clinical research forms (CRFs) into the study database. These data were anonymized and stored in a secure central database hosted on Cambridge University Hospitals NHS Foundation Trust servers. Several data management issues should be noted as the data collection process may introduce errors or biases. Errors during the baseline interview (e.g., failure to ask questions or record a response) or lack of information in the case notes/electronic records may

contribute to missing data. As many patients are of advanced age, recall bias may also introduce discrepancies (e.g., answers to history of heaving drinking or smoking).

Data preparation and variable construction

Processing of baseline clinical and epidemiological data

The raw and fully anonymized baseline data for OCCAMS (R Data file format, .Rdata) and BEST2 (comma-separated values file format, .csv) were exported to my university-furnished computing device in June 2022. The files containing the data for OCCAMS and BEST2 were collated, processed, and screened for completeness, accuracy, and consistency. Data were cleaned, removing or correcting any inconsistencies, inaccuracies, or implausible values. All pragmatic strategies to minimize missing data were implemented. The cleaned dataset was then carefully checked against the raw data to ensure quality data pre-processing. The datasets and data-cleaning code were saved as plain-text files and tracked and managed using version control software (Git/Subversion). All data processing was carried out using R Version 4.2.3 on macOS Ventura 13.3.1.

The following common methodology was used to clean data for both OCCAMS and BEST2 studies. Due to the inclusion criteria, age at diagnosis and sex were complete. Ethnicity was recoded into white or other as there were too few observations in other ethnicity codes which is not unexpected for the BE/EAC patient population.

The age at diagnosis for EAC and BE cases and age at recruitment for reflux controls, as well as BMI at baseline, were categorized into groups. This was done to create a more meaningful comparison for these measures. Age at diagnosis was categorized into four groups of under 50 years old, 50-59 years old, 60-69 years old and 70 years or older. BMI was calculated using the baseline weight and height (weight in kg/height in m squared) and

92 BMI categories were defined according to standard ranges of underweight (<18.5 kg/m^2),
93 normal (18.5-24.9 kg/m^2), overweight (25-29.9 kg/m^2) and obese (≥30 kg/m^2). Underweight
94 cases were included among normal weight due to very small frequencies in the cohort
95 (<2.5%). The continuous distribution and the grouped frequencies were used in descriptive
96 analyses and only the categorical variables for age and BMI were included in regression
97 models.
98 Cigarette smoking was collapsed into a binary variable with 'former' and 'current'
99 recoded as 'ever' smoker and 'never' remaining as defined. Additionally, if the average
100 number of cigarettes per day was recorded as zero, smoking status was set to 'never' and if
101 it was a non-zero value then smoking was recoded to 'ever'. The number of pack-years was
102 calculated by multiplying the number of packs of cigarettes smoked per day by the number
103 of years of smoking.
104 The self-reported responses for medication use frequency of aspirin, NSAIDs, PPIs,
105 H2RAs and over-the-counter acid (OTC) suppressants included 'Never,' 'No,' 'Past Use,'
106 'Occasional Use,' and 'Current Use.' However, responses such as 'Past Use' and
107 'Occasional Use' are open-ended, so to mitigate this issue, responses were recoded to binary
108 'Ever Use' and 'Never Use'. The duration of medication use was recorded in years, months,
109 weeks and days. The total duration of use (in years) was calculated for each medication type
110 by summing the individual measures. Additionally, if the frequency was set to 'Never' and
111 a non-zero total duration of use was reported, the total was set to null. Conversely, if a non-
112 zero duration was recorded, then the frequency of use was set to 'Ever Use'. Aspirin and
113 NSAID frequency of use were combined into a single variable measuring use of either

114 medication. Similarly, a single variable for the use of any acid suppressant medication was
115 derived using the frequency of use of PPIs, H2RAs and OTC acid suppressants.
116 Alcohol intake was recorded as the number of units of beer, wine and spirits consumed
117 per week. These individual measures were summed into a single continuous variable for the
118 total number of alcoholic drink units consumed per week. Heavy drinking status was self-
119 reported by patients in both studies.
120 Frequency of reflux symptoms was reported as 'Never', 'Sometimes', 'Often', 'Daily'
121 and 'Unknow/sporadic'. Duration of reflux symptoms was harmonized into an ordinal
122 variable with four ranges (Never, 5 years, 5-10 years, > 10 years and unknown/sporadic). A
123 single variable was created that combined all measures related to reflux symptoms and acid-
124 suppressant medication use. This variable is referend to as the "derived heartburn symptoms
125 status" variable (Figure SM2). In addition, a single variable for use of any acid-suppressant
126 medications was derived based only on the acid-suppressant medication use variables
127 (patient on acid suppressant and use or duration of PPIs, OTC acid suppressant medications
128 or H2RAs).
129 For variables that contained responses with undefined free text or numeric ranges instead
130 of a single value, the response was either set to missing or the mid-range was calculated.
131 For example, a free text input of 'undistilled only' for total alcohol unit intake was set to
132 missing and a response of '3-5' cigarettes per day was recalculated to '4' per day. Continuous
133 variables where a negative numeric value was recorded were recoded to missing as per CRF
134 instructions.
135 The UK regions for OCCAMS and BEST2 study centers were determined based on their
136 locations and classified using the International Territorial Level 1 (Office of National

Statistics). Finally, for EAC cases only, combined TNM staging was created according to the UICC/AJCC 7th edition guidelines[1].

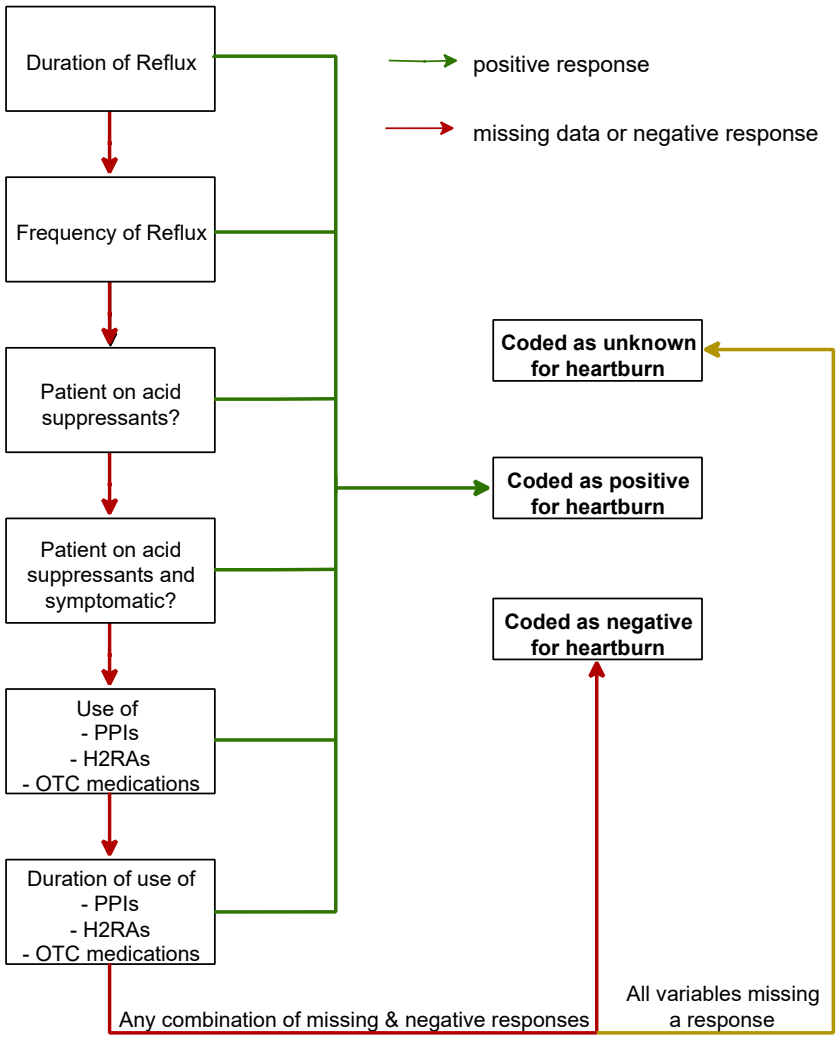

Figure SM1 – Schematic for deriving the heartburn variable using a combination of reflux-related variables. Abbreviations: PPIs, proton pump inhibitors; H2RAs, histamine H2-receptor antagonists; OTC, over the-counter.

Variable selection

Following baseline data cleaning and screening in the OCCAMS cohort and as informed by the results of the literature review, a total of 34 variables across five domains were

deemed relevant and included (Table 1). To select variables for inferential analysis, a purposeful selection process was followed:

1. Unconditional logistic regression was used to obtain univariable ORs and 95% CIs for the association of each variable with BE-ve EAC compared to BE+ve EAC cases.

2. Variables with a p-value <0.25 and missing data <60% overall were pre-selected and included in a multivariable logistic regression model with BE-ve EAC as the outcome compared to BE+ve EAC.

3. Only variables with a p-value <0.05 or those deemed to have epidemiological or clinical importance were selected in the final stage. Directed acyclic graphs were also used to determine which variables should be included as potential confounders.

Table 1 – List of variables included for the analysis from the OCCAMS study. Bold indicates variables which were selected for further analysis using the selection process.

Domain	Variable
Demographics	**Age at diagnosis**
	Sex
	Ethnicity
Risk factor exposures	**BMI at baseline; *kg/m²***
	BMI five years prior to diagnosis; *kg/m²*
	BMI difference (prior to baseline); *kg/m²*
	Cigarette smoking status
	Number of cigarettes smoked per day
	Years of smoking cigarettes
	Number of pack-years of smoking
	Heavy alcohol drinking status
	Units of alcohol intake per week
Anti-inflammatory medications	Aspirin use status
	Years of aspirin use
	NSAID use status
	Years of NSAID use
	Any use of aspirin or NSAID
Reflux symptoms & acid suppressant medications	Frequency of reflux symptoms
	Duration since reflux symptoms began
	Currently taking acid suppressant medications
	Currently symptomatic for reflux while on acid-suppressants
	PPI medication use status
	Years of PPI medication use
	OTC acid suppressant medication use status
	Years of OTC medication use status
	H2RA medication use
	Years of H2RA medication use
	Use of any acid suppressant/reducing medications status
	Derived heartburn symptom status
Clinical factors	Tumor length, *cm*
	Siewert-Stein classification
	Tumor location (resection pathology)
	Tumor growth (T stage pre-op to T stage post-op)
	TNM

Whole-genome sequencing

Strict pathology consensus review was observed for these samples, with a minimum 70% cellularity requirement before inclusion. All tissue samples were snap frozen. For the OCCAMS study, peripheral blood was used as the germline reference, and in cases where this was not possible, a sample of normal squamous epithelium located at least 5 cm away from the lesion or normal duodenal tissue was used instead according to standard practice.

Methods for sample quality control, DNA extraction, and WGS were as previously described[2–4]. Briefly, the 710 EAC and the 205 BE samples sequenced by Illumina, the CRUK Cambridge Institute, and the Wellcome Sanger Institute, underwent WGS to a target depth of 50x. Matched germline samples were sequenced to a target depth of 30x. Reads were then aligned with BWA-MEM to the 1000 Genomes Project version of the GRCh37 human reference genome[5] . Each of the 79 BE sample pools and matched germline samples sequenced by Genomics England were processed in two aliquots to combined target depths of 150x and 75x respectively and reads were aligned to GRCh38. Sequencing quality checks were conducted using the FastQC package (bioinformatics.babraham.ac.uk/projects/fastqc) and PCR and optical duplicates were flagged using Picard MarkDuplicates (broadinstitute.github.io/picard) following alignment.

Single nucleotide and copy number variant calling

For the 710 EAC and 205 BE samples, somatic variants were called using Strelka version 2.0.15[6] and annotated using Variant Effect Predictor (VEP) version 78[7]. Mutation burden was derived from each sample's VEP files by summing the number of SNVs and indels

across the genome. For the additional 79 BE samples sequenced by Genomics England, Strelka version 2.9.4 and VEP version 91 were used, and mutation burden was calculated by taking the average across the aliquot VEP files for each sample. LiftOver was used to convert mutation loci between versions of the human reference genome. Mutations per mega-base were calculated using the length of the reference genome (3,137,454,505 bp). GISTIC2.0 was used to detect recurrently deleted or amplified regions of the genome using raw copy number values obtained from ASCAT version 2.1[8,9] for the 710 EAC and 205 BE samples and from Canvas version 1.38.0.1554 for the 79 BE samples.

3.3 Whole-genome sequencing

Selection and calling of driver genes

Previously reported driver genes in EAC were derived from genes listed in Frankell *et al*. (2019)[3] and genomic regions were identified using Ensembl BioMart[10]. These gene regions were then used to extract alterations from the outputs of VEP and GISTIC2.0. Driver mutation status was determined based on the alteration type (e.g., missense, nonsense or frameshift) using Strelka and VEP. One or more affected copies were deemed as a mutation.

To identify driver genes associated with BE, we predicted SNV mutations using observed/expected mutation ratios calculated by dNdScv. Copy number driver genes were identified by overlapping genes located within peak regions detected by GISTIC2.0 with those in the COSMIC consensus and previously identified EAC driver genes.

Mutational signatures

Mutational signatures discovery within the cohort was carried out using SigProfilerExtractor[11] on 997 samples as previously described[2]. The optimal signature configuration was determined by selecting from a range of signature combinations (from 5 to 17) based on the highest stability and lowest Frobenius reconstruction error for a signature combination. The optimal configuration was composed of 14 signatures, and its validity was confirmed by independent analysis using Bayesian methodology from Sigminer[12]. Subsequently, deconstructSigs[13] was employed to deduce the mutational contributions of these processes to each sample across the entire cohort presented here.

Copy number, whole-genome duplication and aneuploidy

An amplification is defined as a ploidy-adjusted copy number greater than 2, and a deletion is defined when the copy number value is 0. The percentage of aberrant genome is calculated as the proportion of the genome, excluding sex chromosomes, where the rounded copy number does not equal the rounded ploidy. The fraction of loss of heterozygosity (LOH) is defined as the percentage of the genome where the minor allele frequency is less than 0.5, relative to the entire genome excluding sex chromosomes. Raw copy number values from ASCAT and the PCAWG-11 consensus purity pipeline (github.com/PCAWG-11) were used to determine samples with whole-genome duplication based on tumor ploidy and the extent of loss of heterozygosity[14] . Per sample ploidy and purity were also inferred using this method.

Identification and classification of amplicon events

Copy number segments were called using CNVkit version 0.9.8 and regions of amplifications of size 50kb, copy number > 4.5 were used as input for the identification of amplified regions and reconstructed using Amplicon Architect[15,16].

The classification of amplicons into ecDNA and BFB events was done using Amplicon Classifier (github.com/jluebeck/AmpliconClassifier).

Multiregional mutational lineage tracing

Sample preparation, whole exome sequencing (WES), variant calling and mutational clustering were performed as previously described[17] . ClonEvol (v0.99.11) was used to create the phylogenetic trees, with clusters containing fewer than five mutations excluded. Indels and copy number driver events were added to the trees post hoc.

Statistical analysis

Logistic regression

Since the outcome in each comparison was dichotomous and the association of covariates was non-linear, logistic regression was appropriate. Multinomial logistic regression was also considered; however, it was determined that binary logistic regression would be more appropriate due to its simpler interpretation. The statistical independence of the outcomes

was assumed based on the absence of repeated events and the binomial distribution of the residual variation. It is rare for this assumption of logistic regression to be violated.

The process of selecting variables for these comparisons was described in 3.2.2 (Table 3.1). To ensure that the assumption of multiplicativity was satisfied, effect measure modification was assessed between BMI and heartburn and aspirin/NSAID use and heartburn. A priori it was known that BMI may modulate heartburn. The latter interaction was tested because the heartburn variable was partly derived from PPI use and NSAIDs may modify the effects of PPI in relation to EAC[18].

For each comparison set, crude and adjusted OR and 95%CI were obtained for the association of the age group, sex, BMI group, cigarette smoking, aspirin/NSAID use, heartburn symptoms and TNM (EAC only) with the outcome phenotype. We performed three separate adjusted analyses per comparison: 1) minimally adjusted for age and sex only, 2) fully adjusted for all covariates and 3) fully adjusted model eliminating heartburn as a covariate. As heartburn may be on the causal pathway, if its elimination as a covariate changed the log odds ratio by more than 10%, then it could be considered a confounder. Missing data were coded as indicator variable.

Three sensitivity analyses were performed to assess the robustness of the estimates obtained using the fully adjusted model for each comparison set. The first sensitivity analysis involved excluding any observations with missing data for the variables in the fully adjusted model, adopting a complete case approach. The second sensitivity analysis utilized estimates derived from multiple imputation data (detailed below). Lastly, a sensitivity analysis was conducted by excluding EAC cases with a history of undergoing BE surveillance.

Missing data

Missing data for baseline characteristics was calculated as a percentage of the total number of cases. The percentage of the recorded values is reported as a fraction of complete cases. For variables dependent on the response to other variables, the missing percentage was calculated as a fraction of cases where the first response variable was available. For example, the proportion of missing data for the duration of cigarette smoking was based on the total number of cases who self-reported current or former cigarette smoking.

Multiple imputation (MI) was performed on the datasets corresponding to each comparison group to assess how missing data might bias the observed associations. Age and sex were complete and therefore not included in MI. BMI group, cigarette smoking, aspirin/NSAID use, heartburn symptoms and TNM (EAC only) were imputed using multiple imputation by chained equations with the appropriate method selected based on the variable type[19].

The missing data were assumed to be missing completely at random, meaning that the probability of a value being missing is not related to other data. This assumption was based on the similar distribution observed for recorded and imputed data. Furthermore, baseline data were collected by numerous research staff, and based on our experience, we assumed that variations in the order of the CRF questions, completeness of each section and other factors may have impacted the quality and accuracy of the data collected. Therefore, we assumed that systematic exclusion of data was unlikely. The number of imputations (m) was set to the percent value of the variable with the highest amount of missing data in each

dataset which was aspirin/NSAID use with approximately 50-60% missing data. The number of iterations (n) was set to 20 as typically recommended[20].

Multiple Correspondence Analysis

To delineate the genomic differences between BE+ve and BE-ve EAC, we employed multiple correspondence analysis (MCA) focusing on mutations in recognized EAC driver genes. Each mutation type was categorized distinctly to enable comprehensive analysis. This analysis was performed using the FactoMineR package in R, a robust tool for multivariate exploratory data analysis. Visualization of the MCA results was accomplished using the fviz_mca_ind function from the factoextra package.

Non-parametric data, transformations and multiple hypothesis testing

Statistical comparisons between groups were performed using either the Kruskal-Wallis test or the Mann-Whitney U test, as indicated by the normality of the data distribution. When applicable, data were log-transformed to ensure normality. The percentage of driver genes between groups were compared with Chi-square testing. In cases where multiple comparisons were made, adjustment for false discovery rate using the Benjamin-Hochberg (BH) procedure or Bonferroni correction were applied.

Computing environment

All analyses were performed using R Version 4.2.3 (R Foundation, Vienna, Austria) on macOS Ventura 13.3.1 with packages ‘rstatix’, ‘mice’, ‘survival’, ‘survminer’ and ‘coxme’.

References

1. Edge, S. B. & Compton, C. C. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. *Ann Surg Oncol* **17**, 1471–1474 (2010).
2. Abbas, S. *et al.* Mutational signature dynamics shaping the evolution of oesophageal adenocarcinoma. *Nat Commun* **14**, (2023).
3. Frankell, A. M. *et al.* The landscape of selection in 551 esophageal adenocarcinomas defines genomic biomarkers for the clinic. *Nature Genetics 2019 51:3* **51**, 506–516 (2019).
4. Katz-Summercorn, A. C. *et al.* Multi-omic cross-sectional cohort study of pre-malignant Barrett's esophagus reveals early structural variation and retrotransposon activity. *Nat Commun* **13**, (2022).
5. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).
6. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* **15**, 591–594 (2018).
7. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, (2016).
8. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, (2011).
9. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* **107**, 16910–16915 (2010).
10. Cunningham, F. *et al.* Ensembl 2022. *Nucleic Acids Res* **50**, D988–D995 (2022).
11. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
12. Wang, S. *et al.* Copy number signature analysis tool and its application in prostate cancer reveals distinct mutational

processes and clinical outcomes. *PLoS Genet* **17**, e1009557 (2021).

13. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. deconstructSigs: Delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol* **17**, 1–11 (2016).
14. Dentro, S. C. *et al.* Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell* **184**, 2239-2254.e39 (2021).
15. Deshpande, V. *et al.* Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. *Nature Communications 2019 10:1* **10**, 1–14 (2019).
16. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput Biol* **12**, e1004873 (2016).
17. Black, E. L. *et al.* Understanding the malignant potential of gastric metaplasia of the oesophagus and its relevance to Barrett's oesophagus surveillance: individual-level data analysis. *Gut* **73**, 729–740 (2024).
18. Jankowski, J. A. Z. *et al.* Esomeprazole and aspirin in Barrett's oesophagus (AspECT): a randomised factorial trial. *Lancet* **392**, 400–408 (2018).
19. van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. *J Stat Softw* **45**, 1–67 (2011).
20. Nguyen, C. D., Carlin, J. B. & Lee, K. J. Model checking in multiple imputation: an overview and case study. *Emerg Themes Epidemiol* **14**, (2017).