Cancer Prevention Trials Unit (CPTU)
Cancer Prevention Group School of Cancer & Pharmaceutical Sciences
King's College London
cptu@kcl.ac.uk
Director: Professor Peter Sasieni

Statistical Analysis Plan (SAP)

Title	BEST3 Statistical Analysis Plan		
Reference	BEST3 SOP 008	**Version Number**	**1.0**
Approval Date	16/12/2019	**Effective Date**	30/12/2019
Review Date	30/12/2022		

	Name	**Position**
Authors	Roberta Maroni Marcel Gehrung Judith Offman	Trial Statistician Statistician Trial Epidemiologist

Approved by	Peter Sasieni Name	Director of CTU/Senior Statistician Position
	 Signature	16/12/2019 Date

If this SAP has been printed or saved electronically, please check Sharepoint to ensure this version is the most up-to-date - https://emckclac.sharepoint.com/sites/LSMcpg

Version	Date Approved	Reason for Change	Author
1.0	XX/XX/2019	NA	

Table of Contents

Abbreviations

AE	Adverse Event
BE	Barrett's oEsophagus
BMI	Body Mass Index
BOC	Benign Oesophageal Condition
CLR	CLuster Randomised
CRF	Case Report Form
DMC	Data Monitoring Committee
EAC	Oesophageal AdenoCarcinoma
GI	GastroIntestinal
GP	General Practitioner
H2RA	Histamine-2 Receptor Antagonists
IM	Intestinal Metaplasia
IQR	Interquartile Range
ITT	Intention-To-Treat
MHRA	Medicines and Healthcare products Regulatory Agency
NHS	National Health Service
NPV	Negative Predictive Value
OGD	Oesophago-Gastro-Duodenoscopy
PLR	Patient-Level Randomised
PPV	Positive Predictive Value
RCT	Randomised Controlled Trial
SAE	Serious Adverse Event
SAP	Statistical Analysis Plan
STAI	Spielberger State-Trait Anxiety Inventory
TFF3	Trefoil Factor 3
VIF	Variance Inflation Factor

1. Purpose and objective

This document contains the **Statistical Analysis Plan** (SAP) for the BEST3 study. The SAP is required by the National Institute of Health to improve reproducibility, transparency and validity of clinical trials.

The table of contents of this SAP follows the one recommended in SOP Barts CTU GEN ST 01 “Statistical Analysis Plan”, version 4.0.

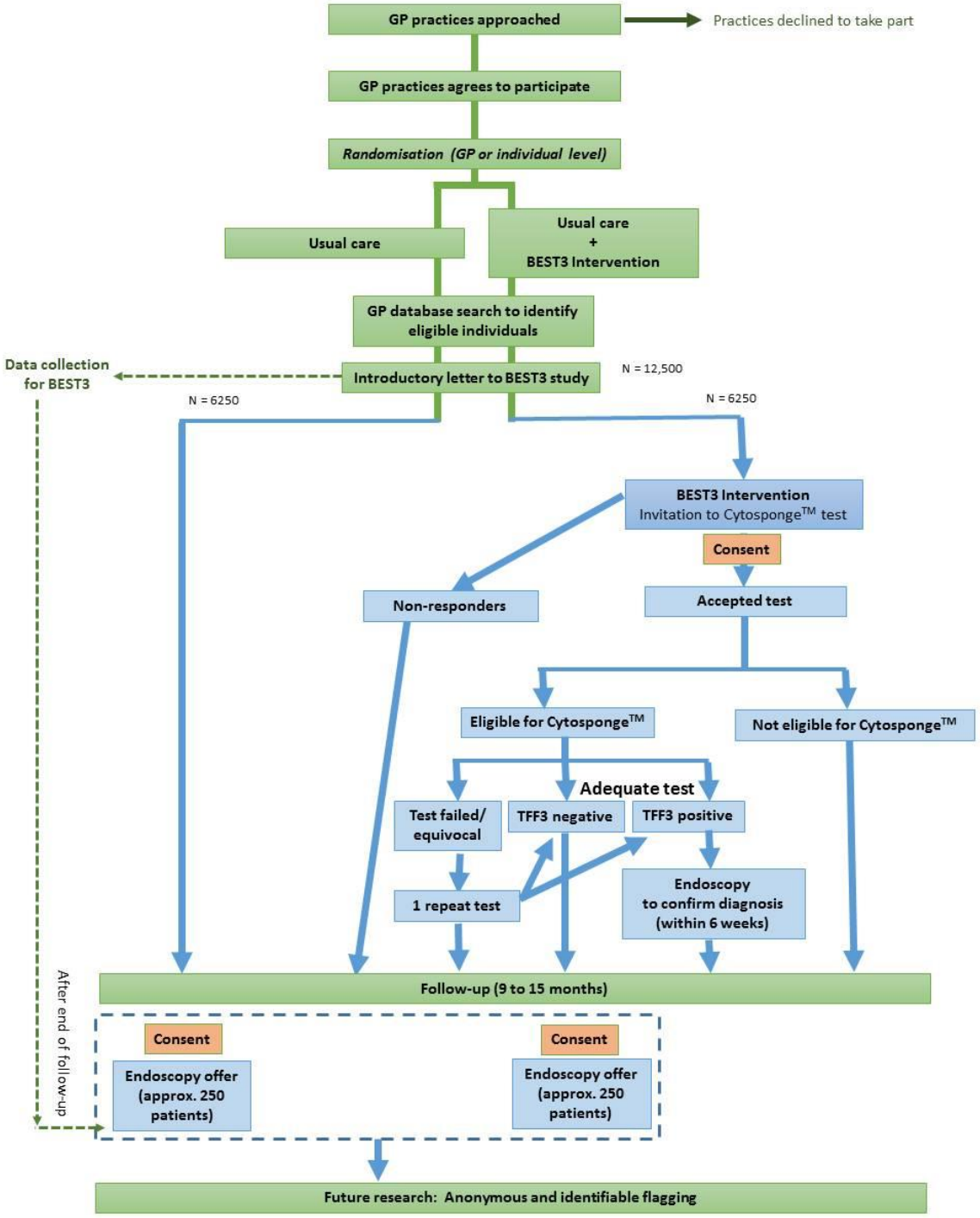
2. Study objectives and design

The BEST3 study is a 1:1 randomised controlled trial (RCT) where consented patients are either recruited to usual care or to receive an invitation to the **Cytosponge™-Trefoil Factor 3** (TFF3) test, a novel non-endoscopic device whose purpose is to detect Barrett’s oEsophagus (BE), a pre-cancerous lesion of oesophageal cancer.

Subjects recruited in the BEST3 study are either cluster randomised (CLR), i.e. General Practitioner (GP) practices are the units of randomisation, or ‘patient-level’ randomised (PLR), i.e. patients are individually randomised to either the intervention or the control arm. This second type of randomisation was added to the study at a later time because of concerns in low recruitment numbers and it was allowed by the fact that some initial conditions of the study were not holding anymore (e.g. GPs were initially expected to recruit patients on an individual basis as they presented, but this was later substituted with automated searches in practice records). Subjects recruited in a PLR fashion confer greater power to the study. For more information on this amendment, see:

G:\EMS\CPTU\BEST3\Section 6 APPROVALS AND AUTHORISATIONS\8. Amendments\Amendment 6

Figure 1. Trial flowchart



2.1 Primary endpoint

The main aim of the trial is to compare the numbers of histologically **confirmed BE diagnoses** between the two study arms. This will confirm whether the Cytosponge™ detects more BE than the current practice (i.e. GP referring a patient to endoscopy).

2.2 Secondary endpoints

To be assessed using data from the intervention arm *only*:

1) Diagnostic accuracy of Cytosponge™
2) Performance of Cytosponge™ in detecting severity of BE
3) Performance of Cytosponge™ in detecting intestinal metaplasia (IM) of the gastric cardia: based on IM detection by endoscopy of Cytosponge™-positive patients without BE
4) Performance of Cytosponge™ in detecting BE or IM of the gastric cardia
5) Performance of Cytosponge™ in detecting oesophageal adenocarcinoma (EAC) and gastric cancer
6) Sampling adequacy
7) Endoscopy referral rates for adequate test results and successful Cytosponge™ swallows
8) Patient acceptability of Cytosponge™: willingness to have the test and to have a repeat procedure, number of patients who fail to swallow, cancer worry, long-term emotional or physical harm, perceived risk of EAC, test experience.
9) Physician/nurse acceptability of Cytosponge™
10) Safety of Cytosponge™
11) Performance of repeat Cytosponge™ test

To be assessed using data from the usual care or *both* study arms:

12) Number of BE diagnoses missed in the current management
13) Number of undiagnosed BE in the general population vs in the group who received Cytosponge™
14) Prevalence and incidence of benign oesophageal conditions (BOCs)
15) Acceptability of endoscopy
16) Number of BE diagnoses for patients with a negative Cytosponge™ result
17) Quality control of endoscopic and pathology results

Epidemiological and longer-term objectives (for up to 10 years):

18) Prevalence and incidence of BE during the study
19) Prevalence and incidence of BE with dysplasia during the study
20) Prevalence and incidence of EAC during the study
21) Prevalence and incidence of IM and related gastric cancers during the study
22) Prevalence and incidence of BE up to 10 years after the study
23) Prevalence and incidence of BE with dysplasia up to 10 years after the study

24) Prevalence and incidence of EAC up to 10 years after the study
25) Prevalence and incidence of IM and related gastric cancers up to 10 years after the study
26) Research and development: Genetic and biochemical risk factors for disease progression (germline and somatic variants and other biomarkers) including targeted, exome level and whole genome sequencing

Health economics analysis:

27) To undertake modelling to predict the reduction in EAC-related mortality from this strategy (short and long term)
28) Cost of the Cytosponge™ test versus usual care
29) Cost-effectiveness of the Cytosponge™ versus usual care

2.3 Assessment of objectives

Assessment of primary endpoint

To assess the primary endpoint, research endoscopy findings (after the end of the follow-up period) will not be taken into account.

Because of the nature of the study, only participants who accepted the invitation to the Cytosponge™ test gave consent to the use of their data for study purposes. In order to collect primary endpoint data in consented Cytosponge™ patients, non-responders of the invitation and usual care participants, at least one of the following different data sources will be used:

1) ***Automated coded search*** for BEs, OGDs, EACs or upper gastrointestinal (GI) referrals: carried out on the GP electronic clinical records for all participants in all sites.
2) ***Manual case-note review*** of the GP clinical record for all participants identified by the coded search plus a number of other participants depending on the capacity of the sites: some practices will include all participants, some will include no additional participants and others will perform a manual review on a sample of participants chosen at random by the Trial statistician.
3) ***National Health Service (NHS) number linkage*** with hospital records for a limited number of sites, depending on the availability of relevant hospital information systems that routinely capture relevant diagnoses and geographical proximity of site in relation to its linked hospital.
   Note: NHS number linkage will not be performed for all hospitals. For instance, in London, this step may be forgone because of the high population density and the consequent difficulty in determining the hospital catchment of sites.

Automated coded searches will be run by the study sites. The central study team does not provide the practices with the codes as different codes are in use in different practices. Following this step, the sites will fill in a spreadsheet prepared by the Trial team asking for the following information by sex and 10-year age range (50-59, 60-69, 70-79, 80-89, 90+): type of upper GI referrals and investigations; upper GI diagnoses during follow-up; treatment for BE. PLR sites will be asked to fill in two different spreadsheets, one for their usual care and one for their intervention patients. They will also indicate in the online database how many endoscopy referrals, BEs and adenocarcinomas (and the affected patients) they will have found in their records.

Manual case-note reviews consist of checking a patient's record closely and filling in an electronic database form with information on: length of follow-up, sex, age range, body mass index (BMI) range, smoking history, drinking history, GP consultations, type and length of acid-suppressant medications, Helicobacter pylori course, aspirin course, medication review, symptoms, any diagnosis following Cytosponge™ test, any treatment for BE.

For NHS linkage, the encrypted NHS numbers of all participants will be sent by practices and NHS numbers of all new diagnoses of BE will be sent by participating endoscopy units to Cambridge University (the "Trusted Third Party") for anonymous matching. NHS numbers will be scrambled using a one-way hashing system (to a SHA256 standard), so will be not identifiable to anyone outside the practice. Sites (GP practices and endoscopy units) will receive an Excel macro tool to perform the encryption. Any endoscopic diagnoses found through this method will be checked directly with the study sites. This step has received approvals from the NHS Research Ethics Committee as part of Amendment 6 to the BEST3 Protocol.

Where there is evidence of endoscopy from one of these sources, further details may be sought from clinical records. Study-related endoscopies for participants with positive or other relevant Cytosponge™ results will be used for this purpose.

Note: it is important to record the date up to which the coded search is looking for BE and OGDs/EACs/upper GI referrals. This date should be used as the date of last follow-up for that patient, i.e. any BE diagnosed after that date should not be included in the primary endpoint.

Table 1: A schematic of how Methods 1 and 2 work together.

		SECOND STEP: manual case-note review returns…		
		BE diagnosis	**Negative**	**Not done**
FIRST STEP: automated coded search returns…	**BE diagnosis**	For all patients with a BE diagnosis picked up in the coded search	N/A	N/A
	OGDs/EAC/referral	Some of the patients with an OGD/EAC/referral will have a BE diagnosis when their record is reviewed.	Some of the patients with an OGD/EAC/referral will not have any BE diagnoses when their record is reviewed.	N/A
	Negative	A randomly selected number of patients who have not been identified by the coded search will have their records reviewed, which will show a BE diagnosis.	A randomly selected number of patients who have not been identified by the coded search will have their records reviewed and no BE diagnoses will be found.	Most patients not identified by the coded search will not have their records reviewed manually.

The primary endpoint will be histological confirmed BE (column "BE diagnosis" in Table 1 plus any BEs picked up by the NHS number linkage exercise) only. (The case-note review should make clear where the BE was confirmed on histology or whether it was just the impression on endoscopic that was not confirmed.) Patients with BE identified via any of the three routes (automated search plus manual confirmation, manual search, record linkage) will be considered to have BE. Patients whose records have not been searched by any of the three approaches will be considered to have missing BE status. All other patients will be considered not to have a BE diagnosis. All patients with a BE/OGD/EAC/upper GI referral should have a manual review to confirm the diagnosis; if they have not had a manual review, their BE status will be considered as (partially) missing and handled through multiple imputations (of the outcome of the missing review).

Approximate balance in the number of full case-note reviews performed at random should be ensured between the two arms for practices in the CLR group.

Important! BEs known to the study team but not identified through at least one of the three methods describes above will not count towards the primary endpoint analysis. This is most likely to apply to cases detected following a positive Cytosponge™-TFF3 stain and resulting in an endoscopy performed by a study gastroenterologist. This rule is to ensure that the same data collection methods are used in the two study arms (as trial endoscopies are not available in the usual care arm) and helps avoid any biases. However, as a sensitivity analysis, all BEs known to the study team (prior to the exit research endoscopies) will be included. . For statistics relating only to the Cytosponge™ (e.g. PPV of the TFF3 stain), all cases of BE will be included.

Assessment of secondary endpoints

To be assessed using data from the intervention arm *only* (include all BE regardless of the route of study ascertainment):

1) Diagnostic accuracy of Cytosponge™ according to endoscopic findings only (focus: length of BE)
    - Positive Predictive Value (PPV) = proportion of patients with positive Cytosponge™ test result who have a confirmed BE diagnosis following endoscopy
    - PPV by length of BE detected (length of BE categories: >= C1, >= C2, >= C3)
    - Negative Predictive Value (NPV) = proportion of patients with negative Cytosponge™ test result receiving a research endoscopy after the end of the follow-up period who have a confirmed diagnosis of no BE
2) Performance of Cytosponge™ in detecting severity of BE, i.e. diagnostic accuracy of the test according to endoscopic and pathology findings
    - PPV of the test by severity of BE (except for score = 0)

    Severity of BE for positive Cytosponge™ patients will be scored after biopsy according to the following table:

Score	BE severity
0	Pathology report not available

1	IM of oesophagus on biopsy and endoscopic findings not seen in categories below
2	C1 or C0 up to M3 + IM
3	C2 or more, C0 M4 or more + IM
4	C3 or more
5	Low grade dysplasia (LGD)
6	High grade dysplasia (HGD) or T1a cancer

3) Performance of Cytosponge™ in detecting IM of the gastric cardia:
    - PPV of the test for detection of IM of the gastric cardia in patients without BE, i.e. BE with IM will be excluded
4) Performance of Cytosponge™ in detecting BE or IM of the gastric cardia:
    - PPV of the test for detection of BE or IM of the gastric cardia, i.e. patients from endpoints 1) and 3) combined
5) Performance of Cytosponge™ in detecting EAC and gastric cancer:
    - PPV of the test for detecting EAC or gastric cancer
6) Sampling adequacy (for first test, and first and repeat test combined):

    An *adequate* result is defined as: high-confidence negative (squamous and glandular cells), low-confidence positive (squamous and glandular cells with IM), or high-confidence positive (squamous and glandular cells with IM or cellular atypia).
    An *inadequate* result is defined as: processing/technical failure, low-confidence negative (squamous cells only), or equivocal (squamous and glandular cells with equivocal TFF3 staining).
    Sampling adequacy will be reported for the following measures for first test, and then for first and repeat test combined (i.e. not considering the result of the first test for those patients having a repeat test):

    - Inadequacy rate = proportion of Cytosponge™ test results that are deemed insufficient/inadequate due to processing or technical failure, low-confidence negative (squamous cells only) or equivocal (squamous and glandular cells with equivocal TFF3 staining), to be reported with number of.
    - Number and proportion of Cytosponge™ test results deemed insufficient/inadequate due to technical failure only
    - Number and proportion of Cytosponge™ test results deemed low-confidence negative only
    - Number and proportion of Cytosponge™ test results deemed equivocal only

    The above measures will also be presented for first test only and for first and repeat test combined, excluding those patients receiving an inadequate test result but not attending their repeat test.

7) Endoscopy referral rates for 'adequate' test results and successful Cytosponge™ swallows:
    - Proportion of positive Cytosponge™ patients out of all the ones with an adequate test result, both for first test only, and for first and repeat test combined
    - Proportion of positive Cytosponge™ patients out of all patients swallowing successfully a Cytosponge™ test, both for first test only, and for first and repeat test combined
8) Patient acceptability of Cytosponge™:

    At first test:

    - Number/proportion of patients invited who show interest
    - Number/proportion of patients invited who receive screening phone call and how many of these are not eligible to have the test
    - Willingness to have the test: number/proportion of patients invited who attend appointment
    - Number/proportion of patients who fail to swallow
    - Number of attempts to swallow per patient

    At repeat test (in case of an 'inadequate' result at first test):

    - Number/proportion of patients with an 'inadequate' first test result invited to a repeat test
    - Number/proportion of patients with an 'inadequate' first test result interested or not interested in a repeat test
    - Willingness to have repeat procedure: number/proportion of patients with an 'inadequate' test result who attend a second appointment
    - Number/proportion of patients who fail to swallow
    - Number of attempts to swallow per patient

    All of these figures for first and repeat tests will be presented in a flowchart similar to the one used for the open Data Monitoring Committee (DMC) reports (see Section 10 for reference).

    At baseline (before first test):

    - Cancer worry and long-term emotional or physical harm as measured by STAI-6, a short-form of the state scale of the Spielberger State-Trait Anxiety Inventory
    - Perceived risk of developing EAC: overall and compared to a person of the same age

    At day 7-14:

    - Cancer worry and long-term emotional or physical harm as measured by STAI-6
    - Perceived risk of EAC: overall and compared to a person of the same age
    - Test experience as measured by visual analogue scale (0 = "Completely unacceptable", 10 = "Completely acceptable")

- Test experience as measured by the Inventory To Assess Patient Satisfaction (5-point ordinal scale with 18 items)
- Difference in STAI-6 scores at day 7-14 and baseline

9) Physician/nurse acceptability of Cytosponge™: assessed through qualitative interviews (not discussed in this document).

10) Safety of Cytosponge™:

See Section 8.

11) Performance of repeat Cytosponge™ test:

- Rate of conversion to an 'adequate' result after an 'inadequate' first test result = proportion of repeat tests that have an 'adequate' result, to report with number.
- Number/proportion of repeat tests that have an 'inadequate' result.
- Chances of a repeat test result being TFF3 positive (high or low-confidence) after a low-confidence negative result at the first test.

The measures above will be recalculated at a second stage excluding any patients invited to a repeat test and refusing to attend.

<u>To be assessed using data from the usual care or *both* study arms:</u>

12) Number of BE diagnoses missed in current management:

- Usual care arm: research endoscopy findings

13) Number of undiagnosed BEs in the general population vs in the group who received Cytosponge™:

- Intervention arm: research endoscopy findings
- Usual care arm: research endoscopy findings

14) Acceptability of endoscopy:

- Intervention arm: number attending Cytosponge™ appointment
- Usual care arm: number attending research endoscopies

Proportion of participants in the usual care arm who attend their research endoscopy invitation compared to proportion of participants in the intervention arm who attend their Cytosponge™ invitation.

15) Number of BE diagnoses for patients with a negative Cytosponge™ result:

- Intervention arm: research endoscopy findings

Number of false negatives of the test and false omission rate.

16) Quality control of endoscopic and pathology results:

A central review by the study team of all endoscopic and pathology records of positive Cytosponge™ patients undergoing a confirmatory endoscopy will be performed. Its results will be compared to the endoscopic and pathology results of the study sites by quantifying the number of BE diagnoses missed, the number of any other malignant diagnoses missed and, if relevant, the number of BEs falsely detected. "True" PPVs will be calculated for overall results (first and repeat test combined).

A similar review will look into the results of research endoscopies for usual care and negative Cytosponge™ patients.

Epidemiological and longer-term objectives (for up to 10 years):

To be detailed in the BEST3 Epidemiological Analysis Plan.

Health economics analysis:

This is detailed in the BEST3 Health Economics Analysis Plan and will not be discussed here.

2.4 Level of significance

The level of significance that will be used in all the statistical analyses is 5%, two-sided. However, 95% confidence intervals will be preferred to *p*-values in the final report.

2.5 Sample size

This section combines the initial power calculations with the ones amended after the Milestone 1 review in January 2018, when the study design was changed from CLR only to CLR and PLR. A more detailed explanation on sample size calculations and variable follow-up periods is available here:

G:\EMS\CPTU\BEST3\Section 26 STATISTICS\26.3 Power calculations\Sample size (following Amendment 6)\BEST3_Sample size.pdf

2.5.1 Changes to study design after Milestone 1 review

The BEST3 study was planned as a cluster randomised trial stratified by practice size in order to achieve a reduction in the variance of the estimated treatment effect. This is done by reducing the coefficient of variation of the cluster size within strata. This gives a higher power to the study.

In the initial stage of the study, uptake of the Cytosponge™ was anticipated to be 50%. A lower uptake than anticipated was the main reason why a change in study design was needed – the sample size for a CLR trial with uptake lower than 50% would have been too large to be sustainable for the study. Given the substantial impact to the power of the trial, it was proposed that an additional PLR randomisation design be added to the trial. The reason for the lower than anticipated participation and for why we considered it important to adapt the design to take account of it is explained below.

Originally, the trial was envisaged as recruiting patients presenting to their GPs with (incident) symptoms of reflux. For this reason, we: (1) felt that GP-practice-level randomisation was essential; and (2) assumed that with the personal endorsement of the GP and the fact that patients were essentially asking their GP for help with their reflux, acceptance of the offer of a Cytosponge™ test would be high (50%).

For practical reasons (related to trial delivery), BEST3 has been randomising “prevalent” patients, i.e. practices identify patients who have drug prescription records indicating reflux and invite them (or not, depending on randomisation) all at once. Identification using prescribing history rather than at the presenting appointment has resulted in a much lower uptake than anticipated (approximately 27%). (Patients may have been on reflux medication for many years without recently having had symptoms or consulted with their GP.) We still believe that, were the Cytosponge™ to be offered by GPs to patients presenting with reflux as routine practice, the uptake would be much higher.

To obtain sufficient power without greatly expanding the number of participants and practices required, we adapted the design to allow PLR. Practices already set up or trained to take part in the CLR continued with this randomisation method. Practices engaged at a later date would use individual randomisation.

2.5.2 Sample size calculations

The sample size calculation is based on the following assumptions:

- p_{BE}: BE prevalence in individuals eligible for the study is 4%
- p_E: 10% of patients in the usual care arm will be referred to endoscopy for clinical reasons (after excluding urgent referral)
- $p_{BE|E}$: The prevalence of BE in patients referred to endoscopy in the usual care arm is 6%
- s_C: Cytosponge™ -TFF3 sensitivity is 85%
- s_E: Endoscopy sensitivity is 100%
- u: uptake of the Cytosponge™ test is expected to be 27%

Since only 27% of patients in the Cytosponge™ arm are predicted to have the Cytosponge™ test and patients who do not take up the offer of the test will have the same management as if they were in the usual care arm, we only expect 1.38% of patients in the **intervention arm** to be diagnosed with BE:

$$\%\ \text{BEs in intervention} = u[p_{BE}s_C + (1 - s_C)p_E p_{BE|E}] + (1 - u)p_E p_{BE|E} s_E$$

Note that:

- $up_{BE}s_C$ is the proportion of patients who get a positive Cytosponge$^{\text{TM}}$test result and have BE
- $(1 - u)p_E p_{BE|E} s_E$ is the proportion of patients who do not take up the Cytosponge$^{\text{TM}}$invitation and who are diagnosed with BE after receiving an endoscopy
- $u(1-s_C)p_E p_{BE|E}$ is the proportion of patients who get a negative Cytosponge$^{\text{TM}}$test result despite the fact that they have BE and who then move on to have BE diagnosed by endoscopy

$$\begin{aligned}\%\ \text{BEs in intervention} &= 0.27[0.04 \cdot 0.85 + (1 - 0.85)0.1 \cdot 0.06] + (1 - 0.27)0.1 \cdot 0.06 \cdot 1 \\ &= 1.38\%.\end{aligned}$$

Furthermore, we expect 0.6% of patients in the **usual care arm** to be diagnosed with BE:

$$\%\ \text{BEs in usual care} = p_E p_{BE|E} s_E = 0.1 \cdot 0.06 \cdot 1 = 0.6\%.$$

For a 90% power in an individually randomised trial) comparing 0.6% with 1.38%, we would need 6764 individuals (3382 in each arm). In Stata, this is given by the following code:

```
power twoproportions 0.0138 0.006, power(0.9)
```

In a CLR trial, individuals within the same cluster do not act independently, so it is necessary to randomise more individuals. We can calculate how many people one needs to include in a CLR trial to provide the same amount of information as one person in a PLR trial. That number is called the variance inflation factor (VIF). With a VIF of 3.0, one would need 9000 patients in a CLR trial to have the same power as

3000 in an individually randomised trial. As we increase the number of invitees per cluster (i.e. GP practice), the VIF increases, too. In simple terms, there are diminishing returns.

The sample size calculated above would have initially been multiplied by the **variance inflation factor** (VIF). We estimated the VIF as follows:

$$VIF = 1 + \left(\left(\frac{k-1}{k} CV^2 + 1 \right) mean - 1 \right) ICC$$

where:

- k, the number of practices in the group
- $mean$, the average size of the practices in the group
- CV^2, the square of the coefficient of variation of the number of patients per practice, which was in turn calculated by dividing the standard deviation by the mean
- ICC, the intra-class correlation coefficient

The **intra-class correlation** (of the proportion of patients with BE) was assumed to be 0.025.

The VIF can be calculated overall or by stratum. After the Milestone 1 review, practices in the cluster-randomised group where divided into two groups: practices whose enrolment and size (i.e. recruitment numbers) was confirmed, and practices whose enrolment was expected and whose size was estimated. For both groups, practices were grouped into strata based on their size: 50-65, 66-90, 91-125, 126-175, 176-225. With data available on 9th January, we estimated the VIF in each stratum to be: 2.44, 2.96, 3.59, 4.77, and 5.91, respectively, for the confirmed practices. We also estimated VIFs for the projected practices and anticipated 11,816 patients in total from 100 practices contributing the equivalent of 2924 individually randomised patients (1724 “confirmed”, 1200projected).

The required **sample size** in a PLR setting was calculated above to be **6764**. The equivalent confirmed and projected sizes are to be subtracted to this number, and the result is the number of patients that needed to be recruited in the **PLR group**:

6764 – 1724 – 1200 = **3840**

and the total sample size overall was 15,656 (11,816 + 3,840) participants.

During the trial, recruitment numbers were checked several times and the size of the PLR group was adjusted accordingly.

As of 13th June 2019, **7844** participants were recruited in the **CLR group** (equivalent size: 2120) and **5390** in the **PLR group** after the initial 14-day opt-out period, for a total of 13,234 participants. Note that the CLR practices eventually recruited fewer patients than expected at the time of the Milestone 1 review, and that implied a higher recruitment in the PLR group, whose participants contribute more to the power of the study.

The table below shows a range of possible values for our sample size depending on how we adjust the estimate of our parameters.

	BE prevalence	Cytosponge™ TFF3 sensitivity	Endoscopy in usual care arm (over 12 months)	BE prevalence in patients getting routine endoscopy	Uptake of Cytosponge™	Intra-class correlation coefficient	Sample size if only individual randomisation	Equivalent sample size accounted for by cluster randomisation (11,816 patients in 100 practices)	*Remaining sample size for PLR group*
Our assumptions	4.0%	85%	10%	6%	27%	0.025	6764	2924	***3840***
Varying assumptions	***3.9%***	85%	10%	6%	27%	0.025	7098	2924	***4174***
	4.1%	85%	10%	6%	27%	0.025	6456	2924	***3532***
	4.0%	***84%***	10%	6%	27%	0.025	6894	2924	***3970***
	4.0%	***86%***	10%	6%	27%	0.025	6638	2924	***3714***
	4.0%	85%	***9%***	6%	27%	0.025	5658	2924	***2734***
	4.0%	85%	***11%***	6%	27%	0.025	8176	2924	***5252***
	4.0%	85%	10%	***5%***	27%	0.025	5048	2924	***2124***
	4.0%	85%	10%	***7%***	27%	0.025	9344	2924	***6420***
	4.0%	85%	10%	6%	***26%***	0.025	7190	2924	***4266***
	4.0%	85%	10%	6%	***28%***	0.025	6380	2924	***3456***
	4.0%	85%	10%	6%	27%	***0.024***	6764	**3012**	***3752***
	4.0%	85%	10%	6%	27%	***0.026***	6764	**2840**	***3924***

2.5.3 Lower uptake of the Cytosponge™ invitation

In the updated sample size calculations of Milestone 1, uptake of the Cytosponge™ invitation was expected to be 27%. However, in the data available in June 2019, uptake was closer to 24.0%, which would be equivalent to 84% power according to the same assumptions as in the section above. At the same time, it should be noted that currently we are expecting to recruit 746 patients more than needed by our sample size calculations, so that would raise the power of the study to 87%.

The following table shows how the power of the study varies by keeping the same sample size and changing some of the initial assumptions.

Uptake	Prevalence of BE in eligible population	Prevalence of BE in those referred to endoscopy	Sample size (with 90% power)	Power with n = 6764
27.0%	0.04	0.06	6768	90%
24.0%	0.04	0.06	8260	83%
24.0%	0.05	0.06	5388	95%
24.0%	0.04	0.05	10126	75%
24.0%	0.05	0.05	6246	92%
23.0%	0.04	0.06	8886	81%
22.0%	0.04	0.06	9350	79%
21.0%	0.04	0.06	10126	75%

2.5.4 Variable follow-up periods

Because of delays in the implementation of the change of study design (from CLR to CLR with additional PLR group), it was decided that participants/practices could have

a follow-up longer than 12 months to compensate for participants/practices who will have a follow-up shorter than 12 months due to time constraints.

Different practices will have different follow-ups as long as the **weighted average follow-up** for the study will be greater than 12 months, as shown by the following formula:

$$\text{avg}_{\text{FU}} = \frac{\sum_{i=1}^{P} F_i M_i + \sum_{j=1}^{Q} F_j N_j}{\sum_{i=1}^{P} M_i + \sum_{j=1}^{Q} N_j} \geq 12$$

where F_i, F_j are the different lengths of follow-up (in months) for each different group, M_i are the numbers of patients with different follow-ups in the CLR group in 'equivalent size' terms, N_j are the numbers of patients with different follow-ups in the PLR group, $\sum_{i=1}^{P} M_i$ is the total size (in 'equivalent' terms) of the CLR group, and $\sum_{j=1}^{Q} N_j$ is the total size of the PLR group.

According to our current projections, all practices will have between 8 and 18 months of follow-up. The end date of follow-up will be when a site performs their local coded search; it will therefore be taken from the Coded Search case report form (CRF).

Note: in case a weighted average follow-up of 12 months or more could not be guaranteed, the power of the study would be less than 90%. The following table illustrates the potential loss in power with shorter follow-up:

Total average follow-up (months)	Factor by which to inflate sample size
9	1.24
9.5	1.18
10	1.13
10.5	1.09
11	1.06
11.5	1.03

A detailed explanation on how to calculate the factors by which to inflate the sample size is available here:

G:\EMS\CPTU\BEST3\Section 26 STATISTICS\26.3 Power calculations\Sample size (following Amendment 6)\BEST3_Sample size.pdf

2.5.5 Randomisation algorithm

For the CLR group, practices were randomised to either the intervention or the usual care arm. This was done using a randomisation algorithm that stratified by practice size.

For the PLR group, a single randomisation list was created using block randomisation, allowing for 40 practices of 250 patients each to be randomised, for a total of slightly more than 10,000 potential patients (depending on the size of the last randomisation block). A further step of randomisation was introduced to decide the order in which the practices would be assigned to the list. In January 2009, in order to allow for more practices to be enrolled into the study, a new randomisation list with 10 more practices of 250 patients each was created.

More details on the randomisation algorithms and validations can be found here:

G:\EMS\CPTU\BEST3\Section 9 REGISTRATION AND RANDOMISATION

3. General analysis definitions

3.1 Study periods

Follow-up periods will be **variable** as shown in Section 2.5.4.

3.2 Study populations

3.2.1 Intention-to-treat population

The intention-to-treat (ITT) population will include all participants enrolled into the study. The following participants will be *included* in the ITT analysis:

- Deceased patients:
  BE status at death will be used as their BE status at the end of the trial.
- Study subjects that moved away from a practice:
  We are unlikely to know how long such patients were followed because we only receive aggregate data regarding most patients. For this reason, we will treat them as if they were followed for the same duration as other patients in that practice unless we know when they moved, in which case their follow-up will be treated as censored at that time.
- Participants lost to follow-up due to local information governance restrictions:
  If we have no follow-up on all patients in a particular practice, that practice will effectively be excluded. The fact that this has happened will be noted. Individuals with a Cytosponge™ test will still be included in the intervention-arm only results assuming that BE found as a result of a positive Cytosponge™ will have been recorded. If a patient withdraws consent for their data to be used for research purposes ('Type 2' objection), they will be excluded.

As detailed data are only available for consented patients in the intervention arm, for all patients we assume a follow-up equal to the follow-up of their practice, with the exception of any known participants invited to a research endoscopy before the end of their follow-up period. However, it should be noted that IDs of participants with a BE/OGD/EAC/upper GI referral will be collected for both arms and available for the statistical analysis.

Subjects who opted out (Type 2 objection to their data being used, or explicitly writing to the study team asking for their data to be removed from the database) *after* being entered into the study (i.e. more than 14 days after receiving the trial introductory letter) will be *excluded* from the ITT analysis and any further analyses. The numbers of such patients in each arm will be reported.

Further exploratory epidemiological analyses to study the risk of BE according to baseline/demographics data may be detailed in a separate SAP for epidemiological analyses.

3.2.2 Per-protocol population

The per-protocol analysis will exclude any patient who we know should not have been randomised had the inclusion and exclusion criteria been followed precisely (for example, a patient outside of the age range). It may not be possible to identify individuals who do not meet the eligibility criteria in terms of acid-suppression use (see below).

The per-protocol analysis will also exclude any study subjects in the PLR group randomised to one study arm but erroneously assigned to the other and any participants who moved away before receiving the invitation to the test.

The per-protocol analysis will be performed for the primary endpoint only.

Using the number of case-note reviews available in a practice as the denominator, we will calculate the proportion of patients in a practice that have less than 6 months' worth of acid-suppressant medication prescriptions in the year preceding baseline. This proportion will be used as a threshold in our sensitivity analyses. However, practices recruited at the late stages of the Trial may not perform any additional case-note reviews other than for patients picked up by the local coded search of their clinical information system due to time restrictions. In order to overcome these issues, we plan to perform three different sensitivity analyses:

Sensitivity analysis 1: Exclude the practices we know have more than 20% of patients with less than 6 months' worth of acid-suppressant medication prescriptions as well as those practices for which this information is unknown i.e. there has been fewer than 20 patients whose notes have been reviewed manually.

Sensitivity analysis 2: Exclude only the practices we know have more than 20% of patients with less than 6 months' worth of acid-suppressant medication prescriptions.

Sensitivity analysis 3: Only exclude individual patients known to have had less than 6 months' worth of acid-suppressant medication prescriptions at randomisation.

Sensitivity analyses will be performed for the primary endpoint only.

Further investigations on this will compare the proportion of BEs detected in patients with less than 6 months' worth of acid-suppressant medication prescriptions vs the proportion of BEs detected in patients with more than 6 months' worth of such prescriptions. Because of the nature of the data (aggregated data + individual-level data for consented Cytosponge™ patients and patients who have their case-note reviewed), this exercise will only be performed with the individual-level data available.

3.2.3 Non-compliance corrected (ITT) population

An adjustment for lack of Cytosponge™ use in the intervention arm (non-compliance) will be made following the method detailed in Cuzick et al., which gives an estimate of the effect of the intervention for those who attend the Cytosponge™ test invitation. This will show the causal impact of the Cytosponge™ on BE detection.

BE detection in compliers in the intervention arm will be compared to BE detection in potential compliers in the usual care, i.e. participants who would have received the test if they had been offered it. In order to do that, the proportions of actual or potential compliers/non-compliers in the two arms is assumed to be the same, and so is the proportion of BE detection in actual or potential non-compliers. The proportion of BE detection in potential compliers in the usual care arm will be estimated as:

$$\frac{no.BEs\ in\ usual\ care - no.BEs\ in\ potential\ non\ compliers}{no.potential\ compliers}$$

Data on number of BEs detected in the two study arms will come from the ITT analysis on the primary endpoint.

The relative protection given by the intervention in those who comply is to be estimated as

$$\frac{S_{11}/N_1}{S_0/N_0 - S_{10}/N_1}$$

where $\{S_0, S_{10}, S_{11}\}$ and $\{N_0, N_{10}, N_{11}\}$ are the numbers of BE detected and the numbers of individuals in the usual care arm, in the non-complier population (intervention) and in the complier population (intervention), respectively.

Note: in order to obtain an estimate of the non-compliance corrected effect, data from the CLR group will be treated in the same way as data from the PLR group, i.e. the two datasets will be merged and the effect of clustering in the CLR group will not be considered. However, in order to obtain confidence intervals for the effect, the variance inflation factor for the study as a whole will be applied to the variance of the effect that ignores the clustering. It is noted that this method assumes that the compliance within cluster is independent of the BE prevalence within cluster. New methodology to better adjust for the clustering may be developed.

3.2.4 Safety population

The safety analysis will look at all participants attempting to swallow a Cytosponge™: both the ones producing a Cytosponge™ sample ('successful swallows'), independently from the test result, and the ones not able to swallow a sponge. The endpoints related to the safety of the Cytosponge™, i.e. total number of AEs, number of AEs per participant, number/proportion of participants experiencing an AE, will be analysed on this population.

3.3 Subgroup definitions

Primary endpoint analysis will be carried out separately for the CLR and PLR groups as asked by the Medicines and Healthcare products Regulatory Agency (MHRA) in a communication on 31 May 2018. If the results from both subgroups favour the study intervention, i.e. the proportion of participants with diagnosed BE is greater (regardless of the level of significance) in the invitation to the Cytosponge™ arm than in the usual-care arm, then a combined analysis shall be performed.

3.4 Treatment assignment and treatment groups

Participants in the study were selected by an automated search in GP databases followed by a manual review of their records. All study subjects received an introductory letter to the study, allowing them 14 days to opt out of anonymised data collection. Following this, GP practices (for the CLR group) or participants (for the PLR group) were randomised to either receiving an invitation to the Cytosponge™ test or to usual care.

Non-responders of invitation to the Cytosponge™ were managed as were the patients in the usual care arm.

Participants with a positive Cytosponge™ result were invited to a **confirmatory endoscopy**. Negative Cytosponge™ patients were subsequently managed as were patients in the usual care arm.

Note: A small sample of patients in the usual care arm and negative Cytosponge™ patients were invited to a **research endoscopy** *after* the end of their follow-up period. The result of their research endoscopy is *not* part of the primary endpoint analysis. A handful of patients invited to the test were also invited to have a research

endoscopy at the beginning of the roll-out of the procedure, but invitations to that group were stopped shortly thereafter, and any research endoscopy results in this group of patients will not be taken into account for any analyses.

4. Patient disposition

4.1 Compliance to the Cytosponge™-TFF3 test

Compliance to test (intervention arm) will be defined on two aspects: attendance and successful swallows.

Participants will be provided with two opportunities to successfully swallow the device. A participant will be considered as having had the Cytosponge™ test if he or she has at least one successful swallow. Attenders will include patients who successfully swallow a sponge and those who present at their appointment but are not able to produce a successful swallow.

Study subjects successfully swallowing a Cytosponge™ may produce a sample deemed inadequate because of processing/technical failures or because the test result is considered to be low-confidence negative (squamous cells only) or equivocal (squamous and glandular cells with equivocal TFF3 staining). These patients will be invited to a repeat Cytosponge™ test.

The following numbers and proportions on compliance will feed into the Trial flowchart (see Figure 2):

- responders: overall, to first invitation letter only, to second invitation letter only
- interested and not interested (out of all responders)
- received screening phone call (out of interested): eligible, ineligible
- attenders and non-attenders/withdrawn (out of eligible)
- produced a successful swallow and unable to swallow (out of attenders)
- ‘inadequate’ samples (out of successful swallows) at first attempt and at repeat test: processing/technical failures, low-confidence negative (squamous cells only), equivocal (squamous and glandular cells with equivocal TFF3 staining)
- participants with ‘inadequate’ samples invited for a repeat test, attending the test, producing a successful swallow or unable to swallow
- ‘adequate’ samples (out of successful swallows) at first attempt, at repeat test, overall (i.e. only repeat test counts for participants having two tests) and in total (i.e. all tests counts): negative (squamous and glandular cells), low-confidence positive (squamous and glandular cells with IM), high-confidence positive (squamous and glandular cells with IM and cellular atypia)

Time from first invitation letter to response will be analysed with a Kaplan-Meier estimate, where the event is “responding to the invite” and survival is “not responding to the invite”. However, it should be noted that, in a handful of cases when patients replied very late (> 1 month after invitation) to their test invite, the nurse was not able to offer an appointment (because the study was no longer working in the area) and the patient was marked as a non-responder.

4.2 Compliance to confirmatory endoscopies

Compliance to confirmatory endoscopies will be measured out of all patients receiving a low or high-confidence positive Cytosponge™ test result. We will report on number of attenders and types of diagnoses.

4.3 Compliance to research endoscopies (after end of follow-up)

Compliance to research endoscopies will be measured out of all patients receiving an invite to a research endoscopy. We will report on number of responders, interested/not interested, attenders and types of diagnoses both overall and by study arm.

Note: only participants in the usual care arm and patients receiving a negative Cytosponge™ test result were invited to a research endoscopy. A limited number of 'non-responders', i.e. patients who did not take up their Cytosponge™ invitation, were also invited to research endoscopies at the beginning of the rollout of this part of the trial; their invites were suspended shortly after. Despite the small figures, we will report on number of non-responders invited and attending a research endoscopy in the final statistical report.

5. Demographics and baseline characteristics

The following data will be available at baseline for each GP practice:

Usual care, intervention and PLR practices will send the study team the following baseline data in *aggregated* form (Excel spreadsheet see below):

- Number of participants enrolled by sex and age group
- Drugs administered and dosage

For PLR practices, the trial arm is not included in the aggregated baseline data.

Sex	Age bracket (yrs)					
	50-59	60-69	70-79	80-89	90-99	Total
Female						
Male						
Total						

Demographics and baseline characteristics will be presented by summary statistics. No statistical tests will be performed to compare these between study arms.

Number of sites, number of participants per study arm and average practice size will also be presented.

Patients who take up the Cytosponge™ invitation will have to complete the following CRFs, with the following information available to the Trial Statistician:

- Personal details CRF: sex, year of birth, ethnicity (sensitive data, available to the Statistician via the Trial Senior Research Application Programmer only)
- Allergies CRF
- Baseline Clinical CRF: height, weight, waist/hip circumference, medication for reflux symptoms and dose, gastro-oesophageal reflux disease impact scale questionnaire, heartburn start, H. Pylori, comorbidities
- Baseline Questionnaire CRF: education, smoking history, alcohol intake, risk perception, STAI 6 questionnaire, family history

A number of patients selected at random, both in the usual care and the intervention arm, will have their baseline data and any data regarding a potential upper GI diagnosis and treatment reviewed at the end of their follow-up period. When possible,

sites will review the records of all their patients and fill in a case-note review CRF for each one of them.

A copy of all CRFs is available here:

G:\EMS\CPTU\BEST3\Section 10 CASE REPORT FORM (CRF)\10.1 Current version\BEST3 eCRFs

5.1 Characteristics collected during the study

Participants taking up the Cytosponge™ invitation will also see the following information gathered on them:

- 7-14 day follow-up questionnaire CRF: questions on different elements of the test experience, perceived risk of oesophageal cancer, STAI 6

5.2 End-of-study data

A number of patients selected at random by the Trial Statistician, both in the usual care and the intervention arm, will have their demographics data collected at the end of their follow-up period. These will be the same patients randomly selected to have their baseline data reviewed and a case-note review CRF will be filled in for each one of them.

The primary endpoint data on BE diagnosis will be collected via local coded search + manual case-note review + NHS number linkage as explained in Section 2.3.

5.3 Prior medications and treatments

Acid suppression medication data at baseline is available in aggregated form for all practices. Only medication dose and drug name (not length of treatment) will be available for all individuals.

A number of patients selected at random, both in the usual care and the intervention arm, will have their medication data at baseline reviewed and their medication data at end of study collected at the end of their follow-up period. A case-note review CRF will be filled in for each one of them. For these patients, length of treatment will be available in three monthly categories up to one year and more than one year.

Aggregated data on medication will be compared to medication data gathered during case-note reviews in those practices performing a review of all of their patients.

6. Interim analysis and timing for analysis

6.1 Interim analysis

A Milestone 1 review was planned in January 2018 after six months of opening the first GP site to evaluate the proportion of eligible individuals per surgery (% of population covered), the proportion of participating individuals (% of eligible population), and the Cytosponge™ uptake. This eventually led to a review of the sample size and of the study design (from CLR to CLR and PLR). For more details on this, see Section 2.5.1.

Closed endpoint data on participants who took up the Cytosponge™ invitation were presented at the closed sessions of the DMC meetings of March 2018, October 2018 and October 2019. Reports for the closed session meetings are available here:

G:\EMS\CPTU\BEST3\Section 17 TRIAL COMMITTEES\Data Monitoring Committee (DMC)\Meetings (agenda and minutes)\Reports for closed session

6.2 Time-points for analysis

Only a statistical analysis at the end of the trial is planned.

The last patients were randomised into the trial on 05/04/2019. The coding for the statistical analysis will start in October 2019. The final data lock is expected to happen on 31 January 2020.

7. Efficacy analysis

The statistical analysis will be run using Stata and R. The trial statistician will write the code, which will then be checked by another statistician. The primary analysis of the primary endpoint will be undertaken independently by a second statistician. If the two analyses do not produce identical results, the two statisticians will review their analyses together to reach consensus.

7.1 Method for analysis of endpoints

7.1.1 Analysis of primary endpoint

The power under various assumptions regarding sensitivity, BE prevalence and intra-class correlation will be calculated based on the actual numbers recruited, the uptake observed and the actual duration of follow-up.

Null Hypothesis: The BE detection rate at 12 months (excluding any found on random exit endoscopies) is the same in the intervention arm and the control (usual care) arm.

Alternative hypothesis: The BE detection rate at 12 months is greater in the intervention arm than in the control (usual care) arm.

The CLR and PLR group will be first analysed separately; if the results from the two parts of the study favour the test, a combined analysis will be performed.

Primary endpoint data will be collected according to the methods explained in Section 2.3. This aims at guaranteeing an equal approach to data collection in the two study arms, but implies that, for the primary endpoint analysis, we will only consider BEs that were ascertained systematically through one of the three methods.

Sites have variable follow-ups. The number of person-years of follow-up will be calculated by taking as the end date the date of the local coded search in a practice and as start date the date the first letter of introduction to the study was sent plus 15 days. Follow-up will be considered until whichever is first: diagnosis of BE, the date of the systematic search for BE, the day before a research endoscopy. BE found on research endoscopy will not be counted towards the primary endpoint.

Rates will be calculated out of 1000 person-years. A single rate will be calculated during follow-up in the control arm. Two rates will be calculated in the intervention arm: the rate within four months of randomisation and the rate beyond four months from randomisation. In order to estimate the average rate within 12 months of randomisation in the intervention arm, a weighted average of these two rates will be taken with weights 2:1.

The methods mentioned in the sections below are taken from:

Hayes RJ and Moulton LH. (2009). *Cluster randomised trials*. ed. Boca Raton, FL: Chapman & Hall/CRC, pp. 178-9.

7.1.1.1 Unadjusted analysis

CLR group

The CLR group is stratified by cluster size, i.e. number of participants per practice, as per the categories defined in the Milestone 1 review of the sample size (see Section 2.5.2): 50-65, 66-90, 91-125, 126-175, 176-225.

As a primary analysis, we will run a regression analysis based on individual-level data, followed by a secondary analysis based on cluster-level summaries to ensure that the conclusions are robust.

We will report on number of clusters and patients by stratum and study arm, and on the weighted average follow-up for the CLR group as shown in Section 2.5.4.

Individual-level data

We will first report on cumulative BE detection rate at one year (/ 1000 person-years) by stratum and study arm, and overall, using individual-level data. The one-year rate in the intervention arm will be estimated assuming a constant rate in the first four months and a (possibly different) constant rate thereafter (up to 18 months).

Primary endpoint data will be analysed by a mixed-effects Poisson regression for BE with fixed effects for the treatment and random effects to account for between-cluster variation (i.e. a random effect for the level of BE in each GP practice), with the number of person-years of follow-up as the offset. Additional fixed effect parameters will be included to account for strata (size of clusters in each stratum: 50-65, 66-90, 91-125, 126-175, 176-225).

The resulting 12-month rate ratio will be reported with 95% confidence interval and will be formally tested to see if it is significantly greater than 1.0 (with a two-sided alpha of 0.05).

To fit the Poisson regression random effects model to data, we will model the random effects using a log-gamma distribution.

In Stata, the `mepoisson` command performs a mixed-effects Poisson regression. See Section 11.2.1.1 of Hayes and Moulton for an explanation on the method. There will be two observations (and two durations of follow-up) for each patient: one for the first four months and a second thereafter. There will be a separate treatment effect for each period. The overall treatment effect will be calculated as the weighted mean of the two treatment effects using the Stata command `nlcom`.

Cluster-level data

As a secondary analysis, we will also analyse the data from the cluster-randomised practices using cluster level data. The analysis on cluster-level summaries will follow closely the method explained in Section 12.3.2 of Hayes and Moulton (see also Example 12.3 in the same textbook for a coding example in Stata in the case where the number of clusters (i.e. GP practices) in each stratum is balanced across study arms).

We will first report on mean BE detection rate (/ 1000 person-years) by stratum and study arm, and overall, using cluster-level data. As before, the cumulative rate at 12 months for the intervention arm will be estimated by dividing the follow-up into two periods: the first four months, and the subsequent follow-up (up to a maximum of 18 months).

By stratum: The rate ratio of BE detection for each stratum (approximate number of eligible patients in the practice) will be calculated as the exponential of the difference of the mean log(rates) for BE detection in the two study arms, which is equivalent to the ratio of the geometric means of the rates in the two arms for that stratum. 95% confidence intervals for stratum-specific RRs are calculated according to the method in Section 10.3.2.2 and Example 10.5 of Hayes and Moulton, using the number of clusters minus 2 as the degrees of freedom for the *t* distribution, the number of clusters per study arm, BE detection rates and standard deviations of cluster rates by study arm.

Overall: The overall estimate of the log rate-ratio will then be calculated as a weighted average of the stratum-specific estimates, with weights depending on the number of clusters per study arm and under the assumption that the between-cluster variance in log-rates within each combination of stratum and study arm is constant.

A *stratified t-test* will allow us to test the null hypothesis that the true rate ratio is 1 and to calculate the 95% confidence interval for the RR. The between-cluster variance for this test will be calculated as the residual mean square from the two-way analysis of variance of BE detection rate on stratum and study arm.

Sensitivity analyses will aim at substituting the empirical between-cluster variance with the following predefined values of ICC: 0.025, lower and upper bound of 50% confidence interval of empirical ICC.

Permutation test (on cluster-level summaries)

To check the validity of our inferences, we will use a permutation test. See Sections 6.2.1 and 10.6.3 of Hayes and Moulton on Restricted Randomisation and Permutation Test.

The stratified design of the CLR group implies that restricted randomisation was used in assigning each cluster to its study arm. Let N be the total number of clusters, M be the number of strata and $\{m_i \mid i = 1, \ldots, M\}$ the size of the strata, so that $\sum_{i=1}^{M} m_i = N$. Then, if we require an equal number of clusters in the two study arms within each stratum, the number of possible allocations is:

$$\frac{m_1!}{\left(\frac{m_1}{2}\right)!\left(\frac{m_1}{2}\right)!} \times \ldots \times \frac{m_M!}{\left(\frac{m_M}{2}\right)!\left(\frac{m_M}{2}\right)!}$$

assuming the number of clusters per stratum is even.

According to the strata chosen after the Milestone 1 review and the number and size of practices as of November 2018, this number should be roughly equal to 3×10^{19}, which is too large for the test to be computationally feasible (in a reasonable time). We will instead select a random sample of 5000 permutations. For each permutation, a t-test comparing BE detection rates between study arms will be performed. If the null hypothesis is true, then the observed effect measure can be regarded as having been randomly selected from this permutation distribution.

In Stata, this is done using the `permute` command.

PLR group

We will first report on the number of sites and patients per study arm, and average site size.

We will report on the weighted average follow-up for the PLR group as shown in Section 2.5.4.

Once again, the cumulative rate of BE diagnoses at 12 months will be estimated by dividing the follow-up into two periods: the first four months, and the subsequent follow-up (up to a maximum of 18 months).

A Poisson regression with BE detection rates / 1000 person-years as the outcome, study arm as the exposure and number of person-years as the offset will be run. The resulting rate ratio will be reported with 95% confidence interval.

Combined analysis (CLR + PLR group)

For the purposes of this analysis, the whole dataset will be considered. The same analysis as for the CLR group will be repeated (Poisson regression with random effects on individual-level data, stratified t-test on cluster-level data and permutation test), with the difference that the PLR group will represent a separate stratum of two clusters: one for patients randomised to the intervention and one for patients randomised to usual care. Note that the VIF for this cluster will be equal to 1 as we assume the ICC to be equal to 0: an ICC of 0 implies that there is no clustering so that individuals within the same cluster are no more similar than individuals from different clusters.

A weighted average follow-up will be calculated for the whole dataset as shown in Section 2.5.4. A further estimate of this will be made by considering only 6764 participants (in equivalent size terms) and we will check that this estimate is greater or equal than 12 months.

7.1.1.2 Adjusted analysis

Baseline data on age groups by sex, and length and dosage of acid suppressant medications are only available at practice level. The aggregated nature of the covariate data causes issues for the adjusted analysis both at individual level and at cluster level. For the latter, this is because adjustments for covariates are carried out with a two-stage procedure (see Section 12.3.2 of Hayes and Moulton) that, at first, relies on a regression model with individual-level data. Moreover, the aggregation of baseline data in PLR sites makes it impossible to separate intervention patients from usual care ones. Therefore, any adjusted analyses will not be possible for the primary endpoint.

7.1.1.3 Sensitivity analyses

As mentioned in Section 2.3, the primary endpoint analysis will be reiterated including also actual data on BE diagnoses in the intervention arm deriving from confirmatory trial endoscopies.

A further sensitivity analysis will impute possible additional cases of BE had all three data collection methods been used for all participants.

Moreover, in Section 3.2.2, we explained that three more sensitivity analyses will performed on the per-protocol population to control for the fact that a number of patients have less than 6 months' worth of acid-suppressant medication prescriptions in the year preceding baseline.

7.1.2 Analysis of secondary endpoints

For the sake of simplicity, the cluster design of part of the Trial will be ignored for the analysis of secondary endpoints.

Any endpoints using data on BE diagnoses will rely on *actual* data available from the Trial, i.e. the methods used for data collection for the primary endpoint will not apply.

The analysis of the secondary endpoints will be further detailed in a separate supplementary SAP.

Using data from the intervention arm *only*:

1) Diagnostic accuracy of the Cytosponge™ according to endoscopic findings:

   PPVs will be presented with 95% Clopper-Pearson (exact) confidence intervals overall, by age group/sex, by duration of acid-suppressant medication prescriptions prior to baseline and by number of columnal cells present on the sponge.

   In Stata, these can be calculated using the command `diagt` or `diagti`. In R, the function `epi.tests` should be used.

2) Diagnostic accuracy of the Cytosponge™ test according to endoscopic and pathology findings, i.e. by score of BE severity:

   PPV will be presented with a 95% exact confidence interval.

3) Performance of Cytosponge™ in detecting IM of the gastric cardia:

   PPV will be presented with a 95% exact confidence interval.

4) Performance of Cytosponge™ in detecting BE or IM of the gastric cardia:

   PPV will be presented with a 95% exact confidence interval.

5) Performance of Cytosponge™ in detecting EAC and gastric cancer:

   PPV will be presented with a 95% exact confidence interval;
   Numbers needed to examine (by Cytosponge™) to detect one OAC or one HGDB will also be calculated.

6) Sampling adequacy:

   Inadequacy rate will be presented with a 95% exact confidence interval.

7) Endoscopy referral rates for adequate test results and successful Cytosponge™ swallows:

   The two proportions will be presented with a 95% exact confidence interval.

8) Patient acceptability of Cytosponge™:

   Proportions will be presented with a 95% exact confidence interval.

   Median number of attempts to swallow per patient will be presented with interquartile range (IQR) and range.

   At baseline:

   - Measures will be presented with median, IQR and range

   At day 7-14:

   - Measures will be presented with median, IQR and range
   - Differences in STAI-6 scores at day 7-14 and baseline will be compared with a Wilcoxon signed-rank test.

9) Physician/nurse acceptability of Cytosponge™: qualitative outcome.

10) Safety of Cytosponge™:

    See Section 8.

11) Performance of repeat Cytosponge™ test:

Measures will be presented with a 95% exact confidence interval.

To be assessed using data from the usual care or *both* study arms:

12) Number of BE diagnoses missed in current management:

It will be estimated according to the following method.
Denote by *B* the number of cases of BE found and by *N* the numbers of participants in the denominator:
B_0 and N_0 refer to the numbers in the control arm (excluding the exit research endoscopies)
B_{01} and N_{01} refer to the numbers on the research endoscopies in the control arm
B_{11} and N_{11} refer to the numbers on the research endoscopies in the intervention arm (all Cytosponge™ negative at entry)
B_{10} and N_{10} refer to the numbers who did not have a Cytosponge™ in the intervention arm
N_{12} had a Cytosponge™ in the intervention arm with B_{12} BEs (excluding research endoscopies)
N_{13} had a positive Cytosponge™ and N_{14} had a subsequent endoscopy. B_{14} had BE found via that endoscopy. B_{15} had BE found subsequent to the endoscopy (i.e. after a "negative" endoscopy). B_{16} had BE found following a positive Cytosponge™ despite not having endoscopy as a result of that positive.
The number of BE cases found by usual care is B_0.
Among the N_{01} with a research endoscopy in the control arm, B_{01} cases of BE were missed under the current management. We need to calculate how many were missed in the $N_0 - N_{01}$ participants in the usual care arm without a research endoscopy.
First, consider how many BEs should have been found in the intervention arm had everyone been fully evaluated.
Had everyone with a positive Cytosponge™ had endoscopy, we estimate that $\frac{B_{14}}{N_{14}} N_{13}$ cases would have been found initially and $\frac{B_{15}}{N_{14}} N_{13}$ subsequently. The estimated total number of BE cases in those with a positive Cytosponge™ is:

$$T_1 = \frac{B_{14} + B_{15}}{N_{14}} N_{13} - B_{16}$$

The number of cases missed by Cytosponge™ could be estimated directly:

$$\frac{B_{11}}{N_{11}} (N_{12} - N_{13}).$$

But because N_{11} is (relatively) small, this number will be unstable. Instead, we will use the sensitivity of the Cytosponge™ from BEST2 (80%) in those who did not have a research endoscopy. We then estimate a total of:

$$B_{11} + 0.25(B_{14} + B_{15}) \frac{N_{13}}{N_{14}} \times \frac{N_{12} - N_{13} - N_{11}}{N_{12} - N_{13}}$$

missed cases among those with a negative Cytosponge™. The estimated total number of cases in those with a negative Cytosponge™ is:

$$T_2 = (B_{12} - B_{14} - B_{15} - B_{16}) + B_{11} + 0.25 T_1 \frac{N_{12} - N_{13} - N_{11}}{N_{12} - N_{13}}$$

Next, we need to consider how many cases would have been found in those having a Cytosponge™ had they been in the usual care arm. We

assume that those accepting a Cytosponge™ may not have the same rate as in those that did not accept. By subtraction, we estimate

$$T_3 = \frac{B_0}{N_0}(N_{10} + N_{12}) - B_{10}$$

cases among people accepting a Cytosponge™ (had they not been offered a Cytosponge™). So, among those using a Cytosponge™, BE was increased by the factor: $\frac{T_1+T_2}{T_3}$.

As in the intervention arm, we do not simply scale up from the research endoscopies in the control arm. Rather we combine the cases observed directly among those with a research endoscopy, by the expected number amongst the others using the intervention arm to scale up. The scale factor needed $R_0 = \frac{N_0-N_{01}}{N_{10}+N_{12}}$, i.e. the numbers of people who did not get a research endoscopy in the control arm, divided by the total number in the intervention arm. The total number of cases in the intervention arm is made up of three parts: those observed by research endoscopy plus those in people who would have accepted the Cytosponge™, plus those among people who would not have accepted the Cytosponge™ is offered. The total is estimated as:

$$T_4 = B_{01} + (T_1 + T_2)R + \frac{T_1 + T_2}{T_3} B_{10} R$$

Hence the proportion of BE missed by current management is $\frac{T_4-B_0}{T_4}$.

13) Number of undiagnosed BE in the general population vs in the group who received Cytosponge™:

The number of undiagnosed BEs in the patients who received the Cytosponge™ will be estimated by multiplying the proportion of BEs detected following a research endoscopy in the negative test group by the number of negative test patients.

The number of undiagnosed BEs in the usual care arm will be estimated by multiplying the proportion of BEs following a research endoscopy in the usual care arm by the number of patients in the usual care arm.

The two proportions of undiagnosed BEs will then be calculated out of the total number of patients in each of the two groups and will be compared using a chi-squared test.

14) Acceptability of endoscopy:

The proportion of participants in the usual care arm who attend their research endoscopy invitation will be compared to the proportion of participants in the intervention arm who attend their Cytosponge™ invitation using a chi-squared test.

15) Number of BE diagnoses for patients with a negative Cytosponge™ result:

Number of false negatives of the test arising from research endoscopies will be used to estimate the false omission rate, where we will use as denominator the number of negative Cytosponge™ patients who attend a research endoscopy invitation. The False Omission Rate will be reported with 95% confidence interval.

16) Quality control of endoscopic and pathology results:

For participants swallowing the Cytosponge™ successfully:

- Number (%) of BE diagnoses missed
- Number (%) of any other malignant diagnoses missed
- Number (%) of BEs falsely detected (if any)
- A “true” PPV for BE will be calculated and presented with 95% confidence interval.

For research endoscopies, usual care arm and negative Cytosponge™ patients separately:

- Number (%) of BE diagnoses missed
- Number (%) of any other malignant diagnoses missed
- Number (%) of BEs falsely detected (if any).

7.1.3 Analysis of further subgroups

Because of the nature of the data, we only have individual-level information available for Cytosponge™ patients. Exploratory analyses may be performed by subgroup created using data gathered during the Cytosponge™ appointment, such as age group, gender, BMI, smoking history, etc.

7.2 Covariates

No covariates will be introduced in the primary endpoint analysis because of the type of analysis and the structure of the data available (see Section 7.1.1.2). It should be noted, however, that the primary endpoint analysis will be performed by period (up to 4 months vs from 4 to 18 months).

7.3 Methods for handling missed data and outliers

Any outliers found in the data will be checked with the study sites when possible. Otherwise, they will be substituted by empty fields.

7.3.1 Handling of dropouts

All study subjects received a letter before their follow-up began to inform them about the use of their data within the Trial and to give them the option of opting out of it before 14 days. However, in a handful of cases, participants withdrew consent to the study after the 14-day period (or the practice alerted the trials team late about the objection) and their records were consequently deleted from the Trial database. File notes were filled in for each one of these withdrawals.

As intervention subjects received further letters inviting them to the Cytosponge™ test, it is more likely that they will have withdrawn of the Trial after the 14-day opt-out period in a higher number than usual care patients. For a similar argument, intervention practices in the CLR group, who were more involved in the trial, were more likely to report to the trials team any late opt-outs than usual care practices.

It should also be noted that, in the PLR group, a handful of patients also opted out after being randomised.

Number of dropouts will be reported on, but they will be excluded from any endpoint analysis. However, because of the aggregated (i.e. site-level) nature of the data on age groups and medications, we will not be able to exclude these patients from any summary statistics on these baseline characteristics. A further sensitivity analysis will see dropouts not excluded and treated as participants without BE.

7.3.2 Handling of missing data in active subjects

We do not expect to see any missing data for any primary or secondary endpoints, except for those outcomes linked to patient acceptability questionnaires, for any

outcomes measured in patients who died or moved away during the trial, and for any sites not performing any manual case-note reviews of their records at the end of follow-up.

As we expect very low percentages of missing data, when dealing with missing values for an endpoint analysis, we will exclude individual records accordingly..

8. Safety analysis

8.1 Summary of adverse events

All of the following will be presented by participants producing a successful swallow at first test, participants producing a successful swallow at repeat test, overall (first and repeat test combined, with only AEs from the repeat test contributing for participants who had two tests) and in total (both first and repeat test counted as separate instances).

8.1.1 Number of adverse events

AEs up to 7 days after receiving the test for participants successfully swallowing the Cytosponge™ test will be presented:

- by type: number and distribution
- by severity (severe, moderate, mild): number and distribution
- by study site: number and distribution; median/range by site
- overall: total number

An example of this is available in the DMC report from October 2018:

G:\EMS\CPTU\BEST3\Section 17 TRIAL COMMITTEES\Data Monitoring Committee (DMC)\Meetings (agenda and minutes)\4. DMC meeting - 30 October 2018\BEST3 DMC Report - Open - 30 October 2018.pdf

8.1.2 Number of patients affected by an adverse event

We will report on total number of patients affected by an AE up to 7 days after receiving the test and their proportion over the number of patients who swallowed a Cytosponge™ successfully.

As patients can experience more than one AE, we will also show the median number and range of AEs per participant.

Number and distribution of patients affected by AEs will be presented by site. Median/range by site should also be presented.

8.2 Analysis of adverse events

No statistical analysis of AEs is planned due to the fact that there is no comparison between study arms. However, we may choose to report some of the figures on AEs by subgroup, such as age group, gender, BMI, smoking history, etc.

8.3 Summary of Serious Adverse Events (SAE)

As for the above, SAEs will only be listed for responders of the Cytosponge™. They will be presented individually stating the participant ID, the event narrative, and the relationship with having undertaken the Cytosponge™.

An example of this is available in the DMC report from October 2018:

G:\EMS\CPTU\BEST3\Section 17 TRIAL COMMITTEES\Data Monitoring Committee (DMC)\Meetings (agenda and minutes)\4. DMC meeting - 30 October 2018\BEST3 DMC Report - Open - 30 October 2018.pdf

Number/proportion of SAE (out of all successful swallows) will be reported.

8.4 Analysis of SAE

SAEs are expected to be a rare occurrence in the Trial, so no statistical analysis is planned.

9. Presentation of analysis

Two statisticians will work on the statistical analysis to ensure its reliability: one will write the code, the other one will review it.

9.1 Reporting of results

A statistical report will be prepared, which will follow loosely the following structure:

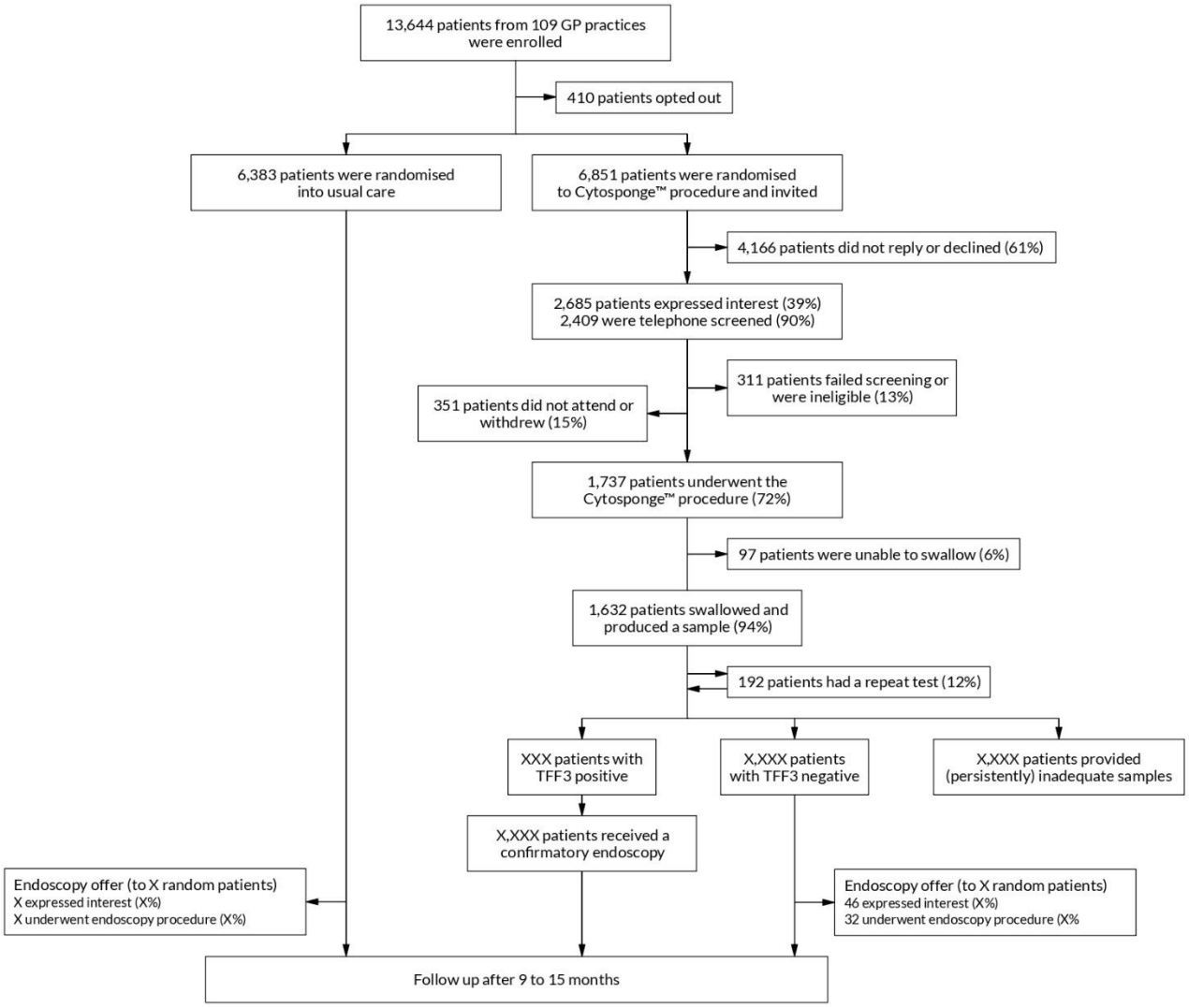
- CONSORT diagram
- Power calculations
- Check on weighted average follow-up
- Patients' demographics summary (for groups/individuals for which these are available)
- Primary endpoint
- Secondary endpoints
- AEs
- Protocol deviations/violations

The CONSORT diagram will be prepared expanding on the Trial flowchart below (temporary figures as of July 2019). The following numbers will be added to the diagram:

- Number of sites who opted out *after* randomisation (CLR group only)
- Patients who opted out *after* randomisation (PLR group only)
- Patients excluded from analysis

Labels for "Enrolment", "Allocation" and "Analysis" will also be added.

Figure 2. Trial flowchart – as of July 2019



9.2 Presentation of results

A statistical report will be prepared. Results will be discussed in a meeting with the study team.

One or more publications will follow.

10. References, related SOPs, web links

SOP Barts CTU GEN ST 01 “Statistical Analysis Plan”, version 4.0

BEST3 Epidemiological Analysis Plan

BEST3 Health Economics Analysis Plan

Randomisation SOP, validation and list: G:\EMS\CPTU\BEST3\Section 9 REGISTRATION AND RANDOMISATION

Cuzick J, Edwards R, Segnan N. Adjusting for non-compliance and contamination in randomized clinical trials. Stat Med. 1997 May 15;16(9):1017-29. Erratum in: Stat Med. 2007 Sep 10;26(20):3821.

Hayes RJ and Moulton LH. (2009). Cluster randomised trials. ed. Boca Raton, FL: Chapman & Hall/CRC, pp. 178-9

StataCorp. 2017. *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LLC

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

RStudio Team (2018). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/

11. Appendices and associated documents

Study protocol: G:\EMS\CPTU\BEST3\Section 4 PROTOCOL

Protocol amendments: G:\EMS\CPTU\BEST3\Section 6 APPROVALS AND AUTHORISATIONS\8. Amendments

DMC meeting reports: G:\EMS\CPTU\BEST3\Section 17 TRIAL COMMITTEES\Data Monitoring Committee (DMC)\Meetings (agenda and minutes)

Sample size calculations: G:\EMS\CPTU\BEST3\Section 26 STATISTICS\26.3 Power calculations\Sample size (following Amendment 6)\BEST3_Sample size.pdf

Communication from MHRA, Your Ref: Amendment 8 (REC Amendment 6), 31 May 2018: G:\EMS\CPTU\BEST3\Section 26 STATISTICS\26.4 Statistical analysis plan\References\Amendment [7] Final Decision Letter.pdf

CRFs: G:\EMS\CPTU\BEST3\Section 10 CASE REPORT FORM (CRF)\10.1 Current version\BEST3 eCRFs