Guideline number	**Guideline 59.006A**	Version	**1.0**
Title	**Cohort Criteria Query**		

1. Procedures

For the purposes of this guideline, cohort creation is the process of generating a list of subjects for inclusion in a research project. Study inclusion / exclusion criteria (see Form 59.001A) are the primary input to the process of cohort production. The output should be a list of subject identifiers that can be used to retrieve from standard datasets, eventually leading to the production of an anonymised extract.

As detailed in SOP 59.001, Data Extracts, a cohort should be created by querying the data for individuals matching the study criteria. Prior to extraction from standard data sets it is sufficient for relevant CHI list in a table designed to hold the cohort criteria. At this point, you are only required to have a list of CHI numbers for the relevant people, until your cohort build is finished.

Create the cohort criteria within separate SQL queries based on the Project Data Specifications. Treating the criteria separately ensures they are easily tested later. Where relevant, quality checked SQL scripts stored in the SQL Library should be used.

1.1. Key Principle

- **Verification of the cohort should be possible throughout the development process.**

Cohort production should lead to a list of subjects that can be used to complete an extraction that supports the answering of a research question. To achieve this end there needs to be some way to verify that the subjects included in the cohort meet the criteria in the project specification. The method outlined here supports this aim. It takes the stance that the cohort identification process should produce artefacts (code and output) throughout and that these can be the subject of testing.

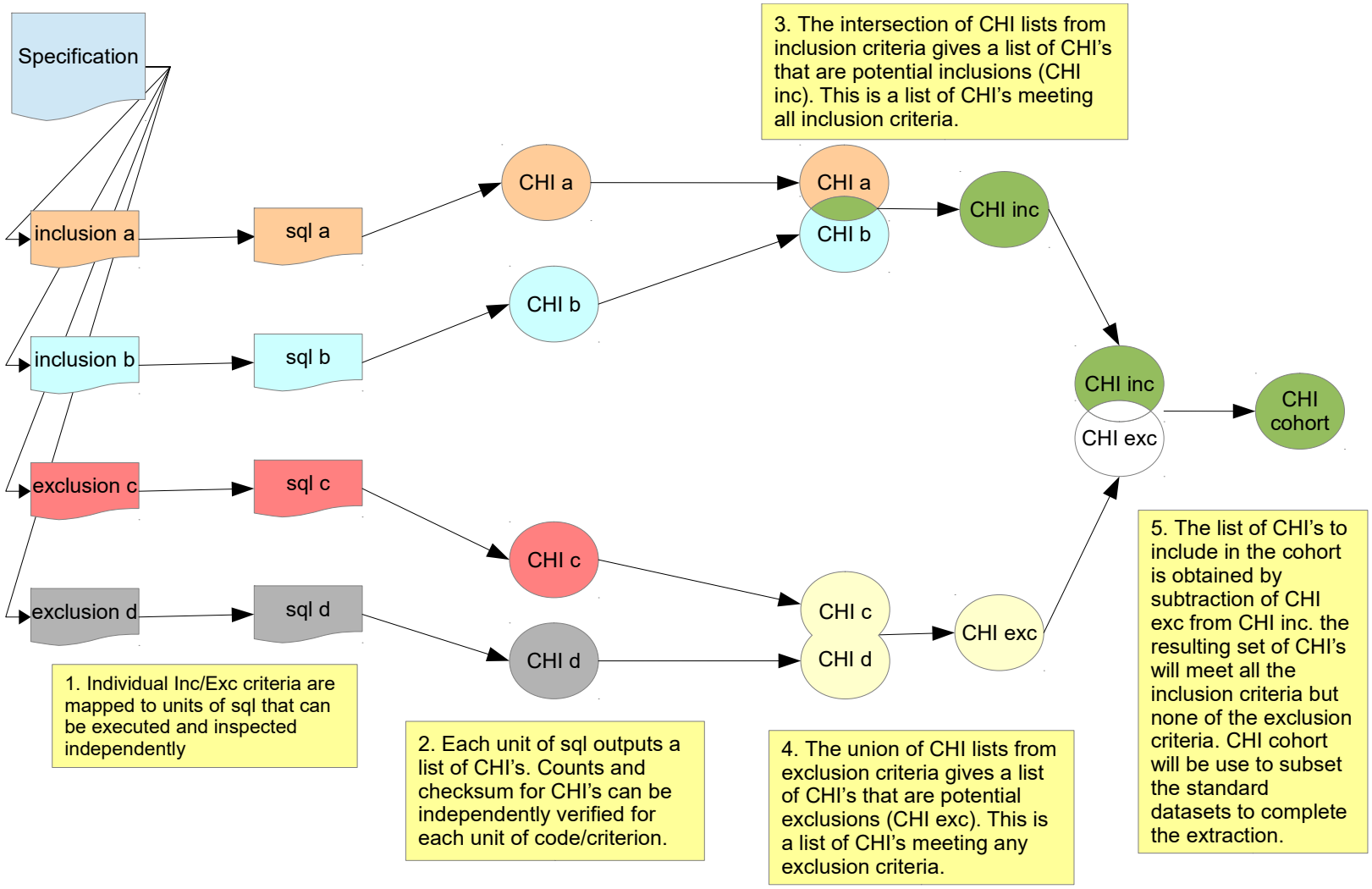
1.2. Algorithm for Cohort Identification

With an agreed statement of study requirements, the process of cohort production can be undertaken in a relatively systematic manner. The following outlines the steps taken by an analyst attempting to generate a cohort.

1. Project specification details independent inclusion and exclusion criteria.
2. For each inclusion/exclusion criterion
    a. Map to a unit of sql for retrieving CHI's
3. For each of the inclusion criterion
    a. Output a list of CHI's
4. For each of the exclusion criterion
    a. Output a list of CHI's
5. Identify the intersection of all outputs from step 3. These are the subjects in the population for whom all inclusion criteria are met.
6. Identify the union of all outputs from step 4. These are the subjects in the population for whom any of the exclusion criteria are met.
7. The cohort is the output from step 5 minus the output from step 6.
8. Document steps 1-7 for review/testing.

These steps are illustrated in figure 1.

Figure 1: Outline of cohort generation process



1.3. Example of Cohort Production

The identification steps outlined above can be readily mapped to a standard pattern applicable to all extracts. For example, the following represents dummy code that might be used for a study - GSH16DUMMY - involving four inclusion criteria and two exclusion criteria. The table names and codes are for illustrative purposes only. However, it should be clear that adding criteria involves little more than adding new code units.

Table 1: Inclusion Criteria:

	Criteria			
ID	**Text**	**Join**	**Filter**	**Date Range**
I1	Has a code for severe asthma in GP LES data		code list[1]	
I2	Has a code for severe asthma in SMR01	**or**	code list	
I3	Age 18 or over	**and**	DOB < 01/04/1996	
I4	Prescribed any of SABA, ICS, ICS/LABA	**and**	code list	01/01/2014 - 13/04/2016

[1]Code list is the read, ICD-10 or local codes identified as applicable from the core dataset.

Table 2: Exclusion Criteria:

	Criteria			
ID	**Text**	**Join**	**Filter**	**Date Range**
E1	Hyper IgE syndrome		code list	
E2	Parasitic Infection	**or**	code list	

Each of the inclusion/exclusion criteria will be specified in natural language and as a list of matched codes and constraining dates within a specified source table in the safehaven TEST environment. A working database will have been created in the TEST environment for use by the safehaven analyst. It is the responsibility of the safehaven analyst to map these requirements to corresponding SQL fragments and embed these in a single batch of code that can be run as a unit when passed to the technical team. In the code below, the CHI numbers generated by each code fragment are output to a holding table (dbo.Cohort_Criteria). The code and the content of this table can then be reviewed and verified as necessary:

```
use Extract_GSH16DUMMY
go


/* CRITERIA TABLE
   A table should be created to hold the CHI's returned
   by each of the inclusion and exclusion criteria. The
   following sections detail the relevant code.
*/

create table Extract_GSH16DUMMY.dbo.Cohort_Criteria ( CriteriaCode varchar(20), CHI
varchar(10))
```

```
/* INCLUSION CRITERIA */

/* I1. Has a code for severe asthma in GP LES data */
insert into Extract_GSH16DUMMY.dbo.Cohort_Criteria
select
	distinct 'I1' as CriteriaCode, CHI
from
	LES_msdi.dbo.[11_GPLES_DefaultDataset] dd
where
	dd.ReadCode like 'H33%'
	or dd.ReadCode = '212G'
-- (7,538 row(s) affected)


/* I2. Has a code for severe asthma in SMR01 */

insert into Extract_GSH16DUMMY.dbo.Cohort_Criteria
select
	distinct 'I2' as CriteriaCode, CHI
from
	[PH_SMR].[dbo].[SH_SMR01_DEFAULTDATASET_13_V]
where
	DIAG1 like 'J4[56]%'
	OR DIAG2 like 'J4[56]%'
	OR DIAG3 like 'J4[56]%'
	OR DIAG4 like 'J4[56]%'
	OR DIAG5 like 'J4[56]%'
	OR DIAG6 like 'J4[56]%'
-- (51,552 row(s) affected)


/* I3. Age 18 or over */
insert into Extract_GSH16DUMMY.dbo.Cohort_Criteria
select
	distinct 'I3' as CriteriaCode, CHINUM
from
	CHI_Download.dbo.CHI
where
	BTHDATE < cast('1996-04-01' as date)
-- (1,268,916 row(s) affected)



/* I4. Prescribed any of SABA, ICS, ICS/LABA */
declare @StartDate as date = '2014-01-01'
declare @EndDate as date = '2016-04-13'
-- (Note: above end date is 4 months and 13 days beyond the protocol end date.dbo.. i.e. the
cohort period)

insert into Extract_GSH16DUMMY.dbo.Cohort_Criteria
select
```

```
        distinct 'I4' as CriteriaCode, Pat_UPI
from
        Pharmacy.dbo.In_Pharmacy
where
        [PI_BNF_Paragraph_Code] IN ('0301011','0301020','0302000','0303020','0301030')
        or PI_Approved_Name = 'Prednisolone'
        and cast(Presc_date as date) between @StartDate and @EndDate
-- (340,301 row(s) affected)



/* EXCLUSIONS */

insert into Extract_GSH16DUMMY.dbo.Cohort_Criteria
select
        distinct 'E1' as CriteriaCode, CHI
from
        [PH_SMR].[dbo].[SH_SMR01_DEFAULTDATASET_13_V]
where
        DIAG1 IN ('D824','C397')
        OR DIAG2 IN ('D824','C397')
        OR DIAG3 IN ('D824','C397')
        OR DIAG4 IN ('D824','C397')
        OR DIAG5 IN ('D824','C397')
        OR DIAG6 IN ('D824','C397')
-- (6 row(s) affected)


/* E2. Parasitic Infection */
insert into Extract_GSH16DUMMY.dbo.Cohort_Criteria
select
        distinct 'E2' as CriteriaCode, CHI
from
        [PH_SMR].[dbo].[SH_SMR01_DEFAULTDATASET_13_V]
where
        DIAG1 LIKE 'H130%'
OR      DIAG2 LIKE 'B7[145789]%'
OR      DIAG3 like 'N220%'
OR      DIAG4 LIKE 'B6[57]%'
OR      DIAG5 LIKE 'B7[145789]%'
OR      DIAG6 like 'N220%'
-- (645 row(s) affected)

SELECT
        x.*
INTO
        Extract_GSH16DUMMY.dbo.Cohort_Distinct
FROM
(
        (
```

```
                select CHI from Dbo.Cohort_Criteria where CriteriaCode = 'I1'
                UNION
                select CHI from Dbo.Cohort_Criteria where CriteriaCode = 'I2'
        )
        INTERSECT
        select CHI from Dbo.Cohort_Criteria where CriteriaCode = 'I3'
        INTERSECT
        select CHI from Dbo.Cohort_Criteria where CriteriaCode = 'I4'
)
-- EXCLUSION BLOCK
EXCEPT
(
        select CHI from Dbo.Cohort_Criteria where CriteriaCode = 'E1'
        UNION
        select CHI from Dbo.Cohort_Criteria where CriteriaCode = 'E2'
)) x


/*
        CHI Dowload
*/


-- insert the 24299 into the Cohort table
-- The windowing function ensures duplicate CHIs are removed, only the latest details
-- being returned from the source CHI dataset.
-- The hashbytes clause is to ensure the records are returned in a random order. This
-- is to simplify assignment of the study ids.
insert into Extract_GSH16DUMMY.dbo.Cohort (CHINUM, BTHDATE, SEX, PCSECTOR, SIMDRANK,
SIMDQUINTILE, SIMDDECILE, SIMDVIGINTILE)
select
        CHINUM,
        BTHDATE,
        SEX,
        PCSECTOR,
        SIMDRANK,
        QUINTILESI,
        DECILESIMD,
        VIGINTILES
from
        (select
                ID,
                CHINUM,
                BTHDATE,
                SEX,
                PCSECTOR,
                SIMDRANK,
                QUINTILESI,
                DECILESIMD,
                VIGINTILES,
                ROW_NUMBER() OVER (PARTITION BY CHINUM ORDER BY ID DESC) as Idx
```

```
        from
                CHI_Download.dbo.CHI chi
                inner join Extract_GSH16DUMMY.dbo.Cohort_Distinct cd
                        on CHI.CHINUM = cd.CHI collate database_default) x
where
        x.Idx = 1
order by
        HASHBYTES('md5', CHINUM+'$'+CONVERT(varchar(50), GETDATE(), 108))
-- (24,299 row(s) affected)


/*
        Obfuscate the DOBs
*/

update
        Extract_GSH16DUMMY.[dbo].[Cohort]
set
        Obfuscated_DOB =
                cast(convert(varchar(4),YEAR(BTHDATE))+'-
'+convert(varchar(2),MONTH(BTHDATE))+'-15' as DATE)
-- (24,299 row(s) affected)




/*
        The dbo.Cohort table is the seed table for extractions from the standard datasets (SMR,
GPLES, etc).
        Essentially all extraction code can follow a similar pattern. For example

        select surrogate_id, field1, field2, ...
        from
        standard_dataset s
        inner join
        cohort c
        on c.chinum = s.chninum
*/
```

To facilitate verification of the cohort extraction steps, the analyst undertaking the extract should record the steps taken to produce the cohort in the Cohort verification document. This should be associated with the version of the Project specification from which the study inclusion criteria were sourced:

Table 3: Cohort verification record – to be reviewed per project specification.

Criterion	Code	Count	Comments
I1	insert into Extract_GSH16DUMMY.dbo.Cohort_Criteria select distinct 'I1' as CriteriaCode, CHI from LES_msdi.dbo.[11_GPLES_DefaultDataset] dd where dd.ReadCode like 'H33%' or dd.ReadCode = '212G' -- (7,538 row(s) affected)	7538	
I2	…		
I3	…		
…			
E2	/* E2. Parasitic Infection */ insert into Extract_GSH16DUMMY.dbo.Cohort_Criteria select distinct 'E2' as CriteriaCode, CHI from [PH_SMR].[dbo].[SH_SMR01_DEFAULTDATASET_13_V] where DIAG1 LIKE 'H130%' OR DIAG2 LIKE 'B7[145789]%' OR DIAG3 like 'N220%' OR DIAG4 LIKE 'B6[57]%' OR DIAG5 LIKE 'B7[145789]%' OR DIAG6 like 'N220%' -- (646 row(s) affected)	646	
Inclusion	(select CHI from Dbo.Cohort_Criteria where CriteriaCode = 'I1' UNION select CHI from Dbo.Cohort_Criteria where CriteriaCode = 'I2') INTERSECT select CHI from Dbo.Cohort_Criteria where CriteriaCode = 'I3' INTERSECT select CHI from Dbo.Cohort_Criteria where CriteriaCode = 'I4'	24326	
Exclusion	select CHI from Dbo.Cohort_Criteria where CriteriaCode = 'E1' UNION select CHI from Dbo.Cohort_Criteria where CriteriaCode = 'E2'	652	
Cohort		24299	

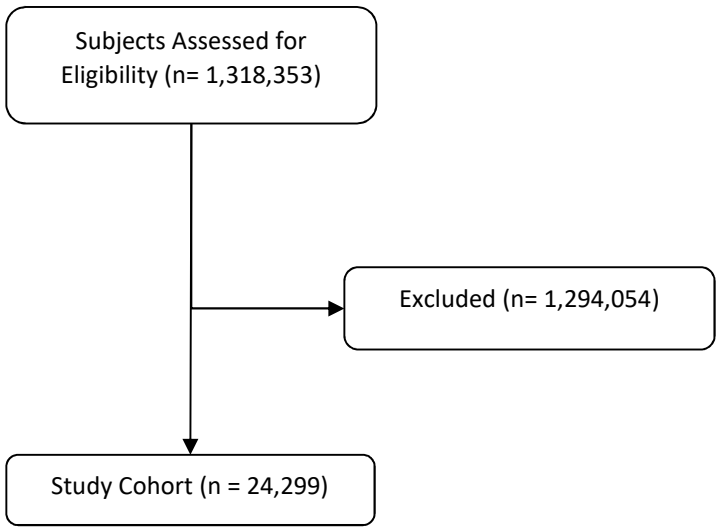
Depending on the risk level identified in the project specification the content of the cohort verification document could be reviewed by a second analyst before the agreed cohort is used as the input for a data extract.

1.4. Alternative Summaries

A key aspect of the pattern described above is that the order that individual code units are executed in is not significant. This is to ensure that the cohort identification process is as robust as possible. However, it may be desirable to report subject counts in a way that reflects a step-wise identification process. Such reports can be produced using the information stored in the Cohort Criteria table, created as part of the cohort generation process.

For example, c
identified for GSH16DUMMY, described in section 1.3. At the highest level, the figures reported at the time of cohort identification could be used (see Figure 2).

Figure 2: High level report of individuals at each stage of study.



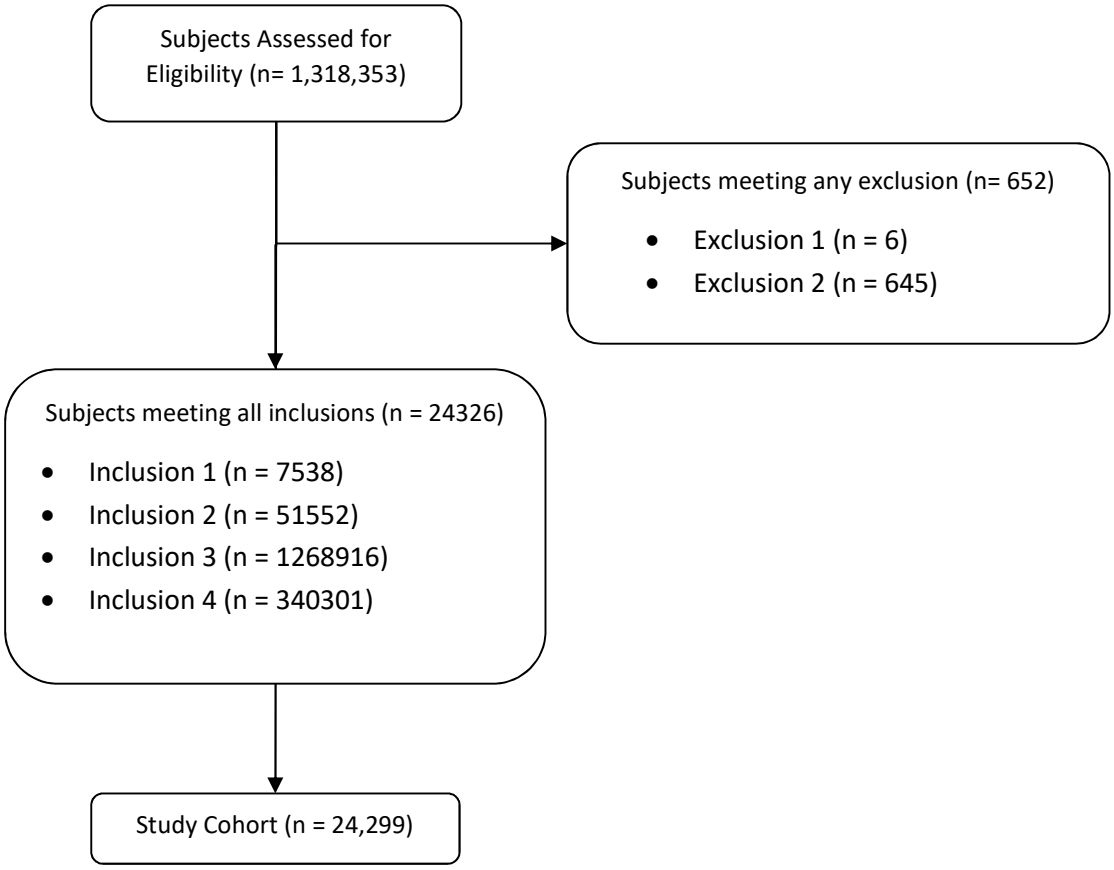
The only additional piece of information required for the above report is a figure for the population originally searched. While the definition could vary (at the extreme, all subjects with data on the platform at a given point in time might be considered the denominator), here it is assumed that a subject meeting any of the inclusion criteria could be a potential subject. This is given by the following code:

```
select count(distinct CHI) from dbo.Cohort_Criteria where CriteriaCode like 'I%'
```

The number excluded is simply the number assessed minus the number in the final cohort.

While this is sufficient for a high level view of the study recruitment process a fuller picture might be required with numbers of individuals at each stage of the study being reported. For example, the effect of individual inclusion / exclusion criteria on the cohort may be required. This can be considered in two ways: the contribution of each criterion can reported across the initial population or the effect of applying each criterion in a stepwise fashion against a gradually reducing sample. The former can be reported using the information that is collected to verify the cohort production and reflects the significance of criterion in the source population (see figure 3). The latter, in contrast represents the impact of applying criteria to an intermediary sample (see figure 4). This too, can be derived from the artefacts produced during cohort identification, with

Figure 3: Report of individuals at each stage of (type 1)



If the impact of applying criteria to successive samples is required, it is important to recognise that the order in which inclusion / exclusion criteria are considered will be of significance with regard to the output produced (this is the primary reason such an approach is not suitable as a testable cohort production process). However, whichever order of execution is deemed appropriate, the information collected as part of the cohort generation process can be used to obtain the required figures. For example, the following code could be used to produce the counts reported in G16 assuming the inclusion and exclusion criteria are to be applied in sequence:

```
-- Subject counts for consort diagram
-- can be obtained by uncommenting sections
-- of code. Note that this code follows
-- the same pattern as that used for cohort
-- identification
select count(distinct x.CHI)
from (
        ---- Uncomment to end of d) when required
        --select distinct CHI
        --from dbo.Cohort_Criteria
        --where Criteria_Code like 'E%'
        --intersect (
        -- a) inclusion 1 or inclusion 2
        (
                select CHI
```

```
                from dbo.Cohort_Criteria
                where CriteriaCode = 'I1'
                UNION
                select CHI
                from dbo.Cohort_Criteria
                where CriteriaCode = 'I2'
        )
        -- end of a)
        ---- b) inclusion 3
        ---- Uncomment to end of b) when required
        --intersect
        --      select CHI
        --      from dbo.Cohort_Criteria
        --      where CriteriaCode = 'I3'
        ---- end of b)
        ---- c) inclusion 4
        ---- Uncomment to end of c) when required
        --intersect
        --      select CHI
        --      from dbo.Cohort_Criteria
        --      where CriteriaCode = 'I4'
        ----end of c)

        --)
        ---- end of d)

) x
```

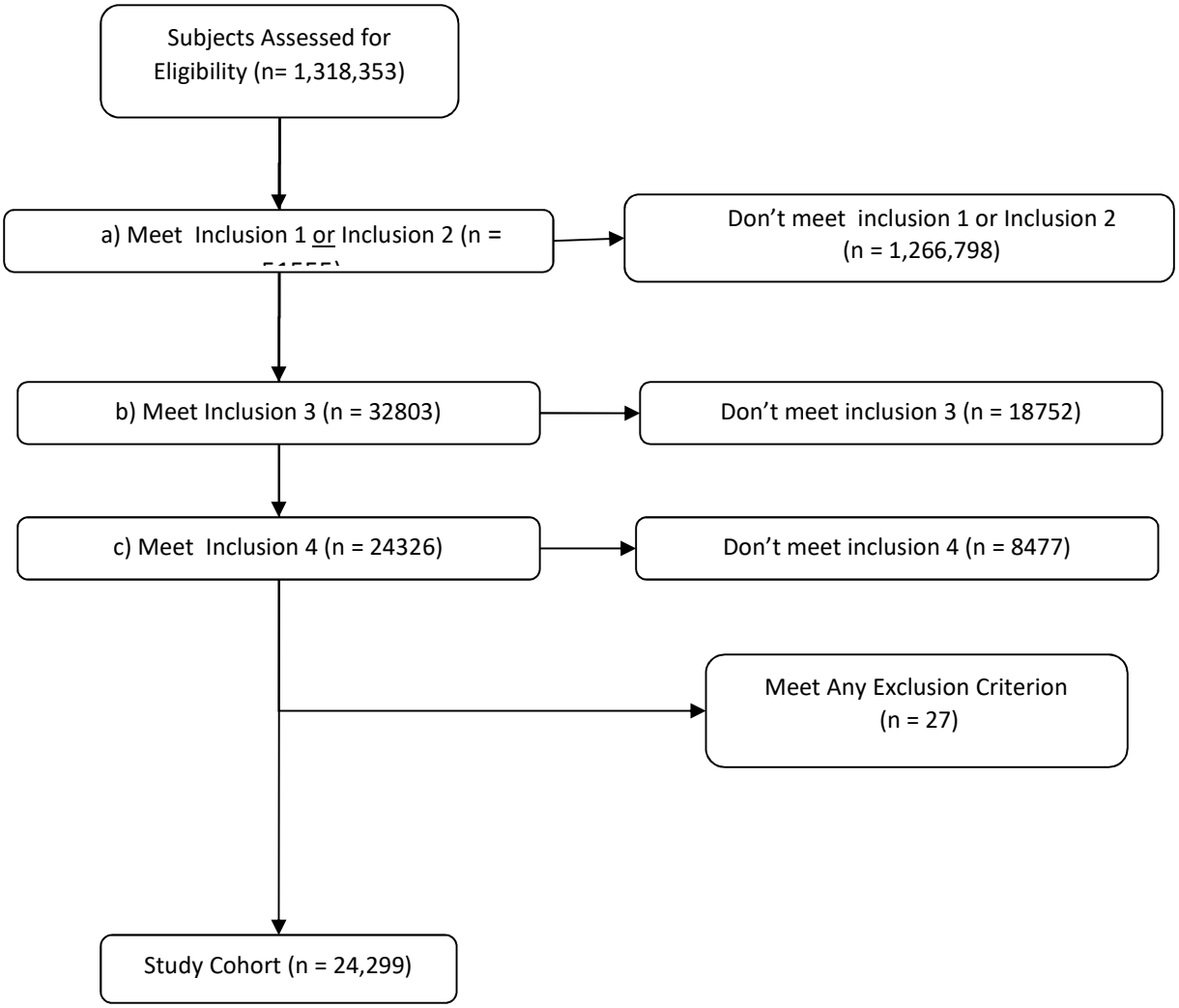


Figure 4: Report of individuals at each stage of identification (type 2)

Although the size of the final study cohort will be unaffected, using this last approach, different counts will be produced for intermediary steps when the order in which inclusion and exclusion criteria are considered is modified. Thus, if such a report is required it is important that the sequence in which criteria are to be applied is considered carefully.

Guideline signatories

Prepared by Signature	Charlie Mayor Date
Approved by Signature	Chloë Cowan Date

Document history

Version	Date	Description
1.0		First Release