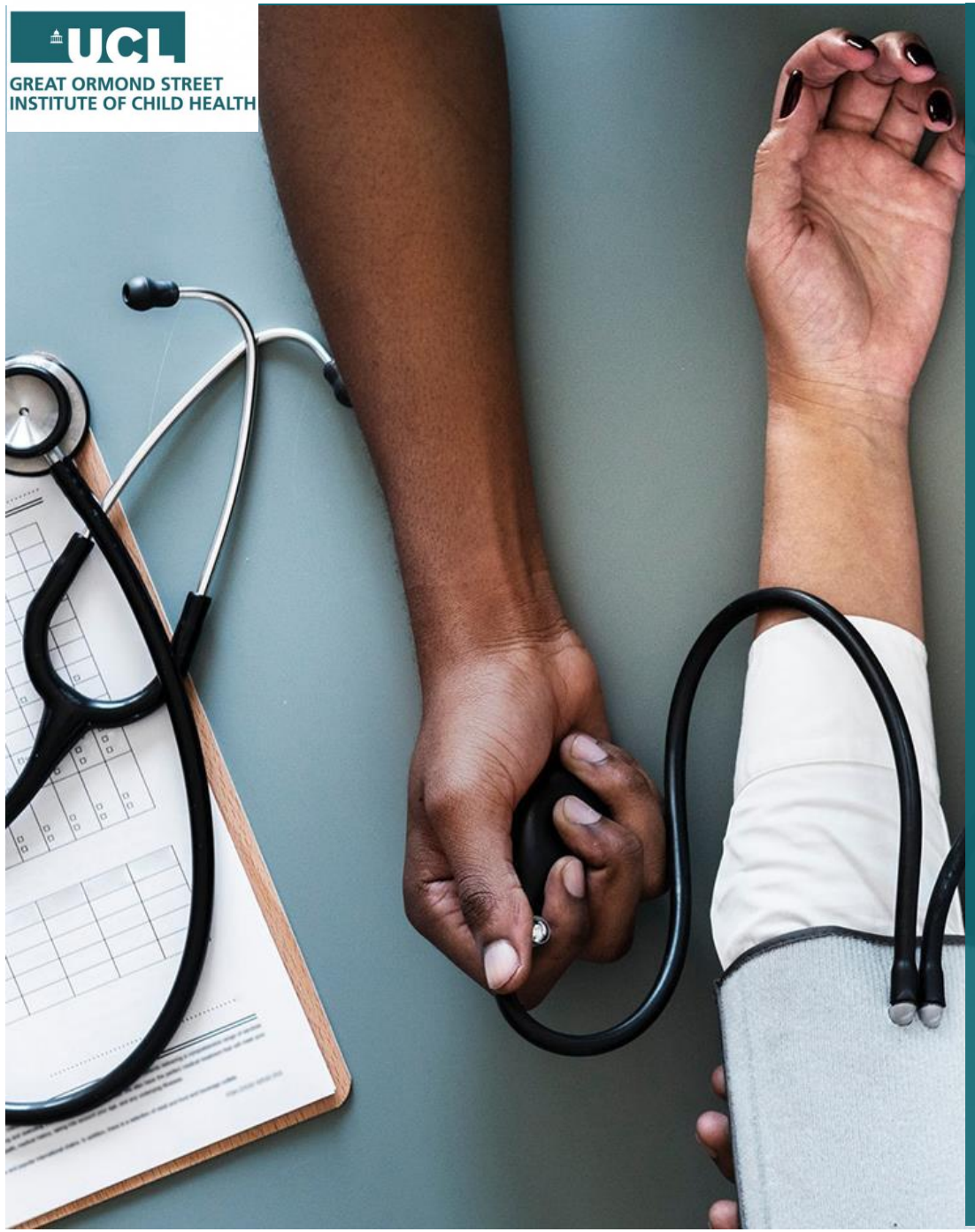


UCL
GREAT ORMOND STREET INSTITUTE OF CHILD HEALTH

acessinehrs.com

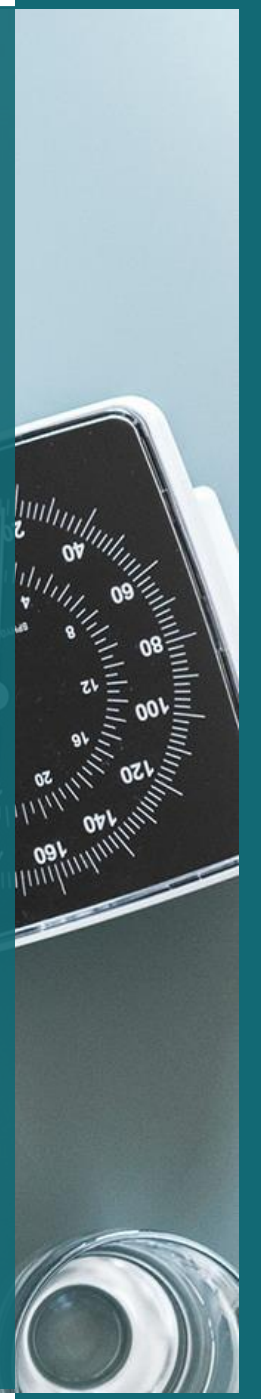
Implementing ACE indicators using R

By Shabeer Syed

Supervised by Prof Ruth Gilbert and Dr Leah Li

UCL Great Ormond Street Institute of Child Health

Population, Policy and Practice Research and Teaching Department



Tutorial aims

Ascertaining ACE indicators in linked data of mothers and children

- **Introduction of data sources and coding systems**

  Read, Snomed CT, Medcodes, prodcodes, ICD codes, entities etc..

- **Data preparation**

  Data cleaning and re-structuring approaches to using multiple data sources

- **How to ascertain ACE indicators**

  Creating a study specific file

  Effectively run code list using vector match - fastMatch

  Spine approach – Merging variables

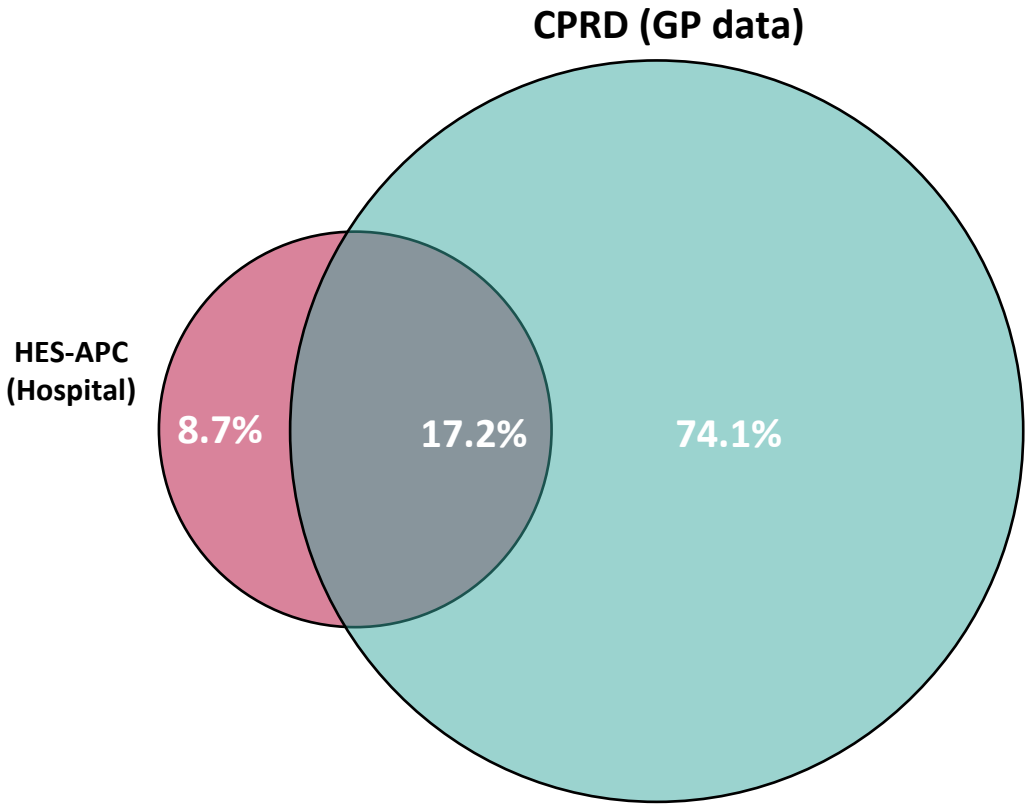
  Hunter, gather, manipulate/make unique, merge cycle

Introduction data sources

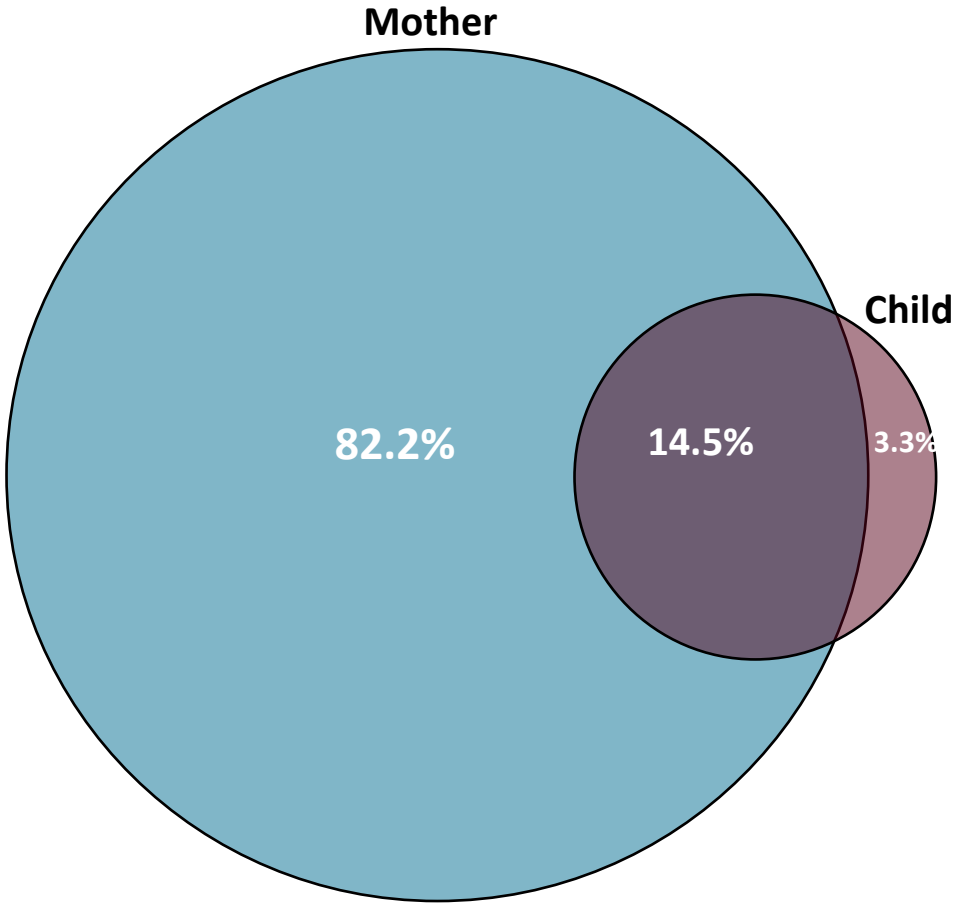
Different coding systems and linked data

ACEs: Data sources needed

Primary care data
Most ACEs are captured in primary care (e.g. GP data, CPRD GOLD, CPRD Aurum).



Mother-child linkage
Most ACEs are captured via mothers record using linked child-mother data.

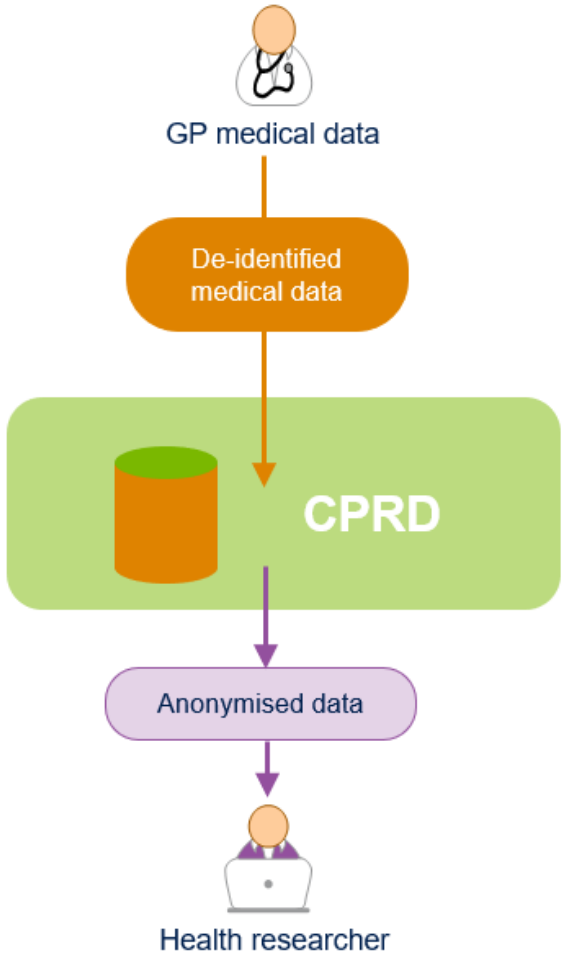


CPRD: Popular GP data source

CPRD GOLD
Most widely used in research. Holds data from software provider named Vision®.

CPRD Aurum
Contains data contributed by practices using EMIS Web® software.

The two systems differs in their data structure and coding. CPRD Aurum database is much larger but more complex, recently introduced, and therefore less used.



- **NHS Data Security and Protection Toolkit**
- **Research Ethics Committee approval**
- **Independent Scientific Advisory Committee**
- **Contractual compliance with CPRD**

GP data representation in England

Open Access Resea

BMJ Open

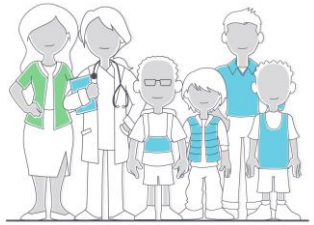
Spatial distribution of clinical computer systems in primary care in England in 2016 and implications for primary care electronic medical record databases: a cross-sectional population study

Evangelos Kontopantelis,[1,2] Richard John Stevens,[3] Peter J Helms,[4] Duncan Edwards,[5] Tim Doran,[6] Darren M Ashcroft[1,7]

Results Of 7526 practices, Egton Medical Information Systems (EMIS) was used in 4199 (56%), SystmOne in 2552 (34%) and Vision in 636 (9%). Great regional variability was observed for all systems, with EMIS having a stronger presence in the West of England, London and the South; SystmOne in the East and some regions in the South; and Vision in London, the South, Greater Manchester and Birmingham.

https://bmjopen.bmj.com/content/8/2/e020738

Mother-baby-link (CPRD)



CPRD Gold (CALIBER Set 16)

- **11.3 million UK patients** (4.4 million active; approx. 7-8% UK pop)
- **695 practices**

↓

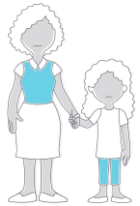
CPRD Mother-Baby-link (1990-2016)

- **771,871** mothers aged 12-50 years
- **1,126,568** linked children
- **695** practices

↓

Eligible for linkage to HES, ONS and IMD (only English GP practices)

- **476,575** mothers
- **697,806** children



CPRD-HES-ONS example data overview

CPRD Spine (one row per patid/wide format)

Mother-baby-link

data

Patient

data

Practices

data

IMD, HES, ONS - Linked data
(Multiple rows per id)

HES Episodes

data

HES Diagnosis

data

HES procedures

data

HES-maternity

data

ONS deaths

data

CPRD data (multiple rows per patid)

Clinical

data

Add. Clinical

data

Referral

data

Immunisation

data

Test

data

Therapy

data

Other linked data (one row per patid)

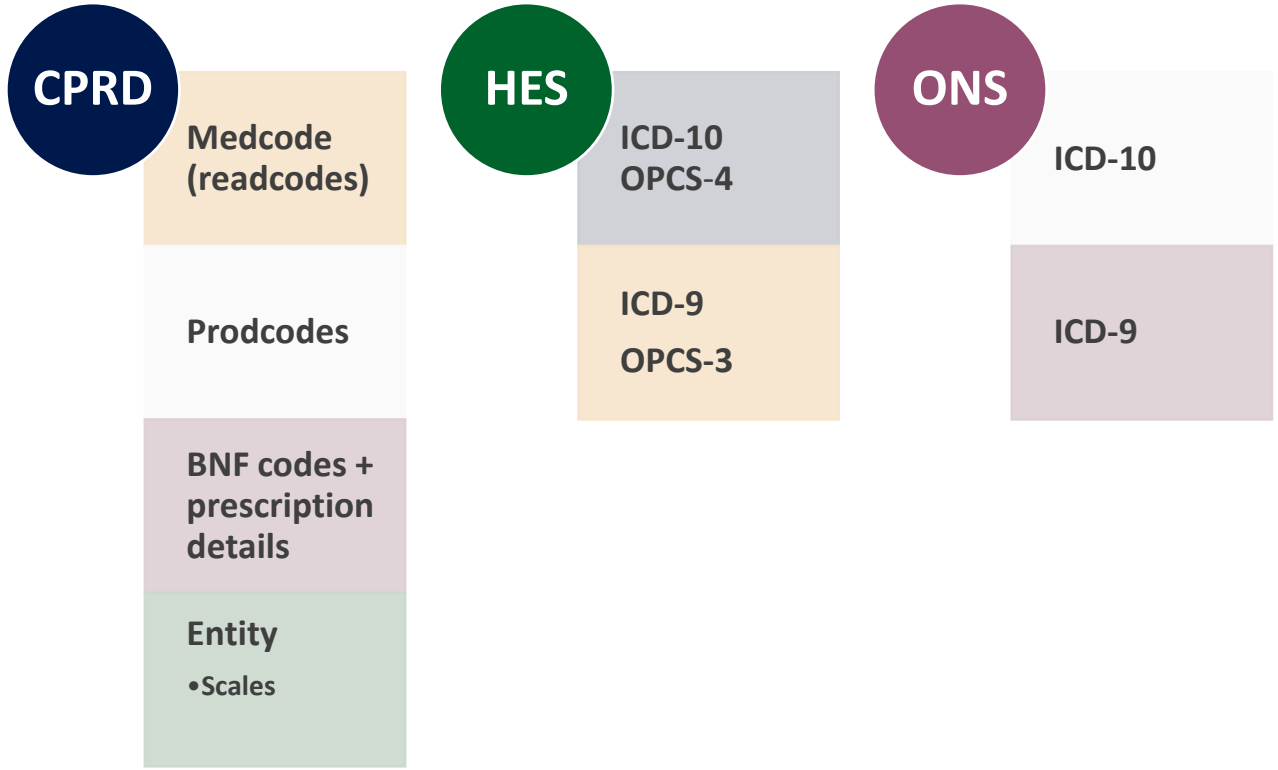
IMD

data

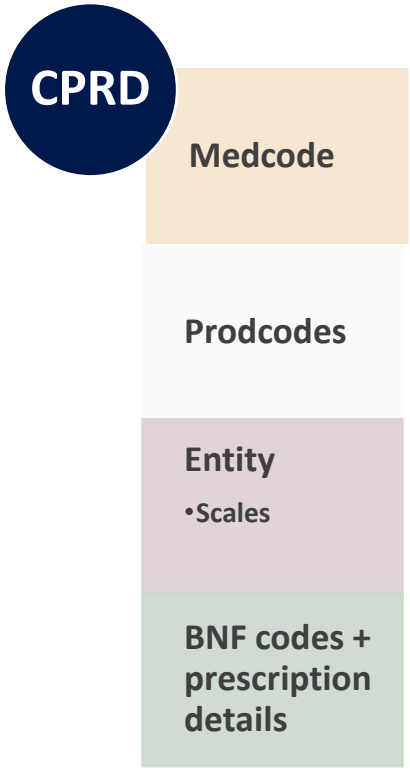
Ethnicity

data

Linked data= different coding systems

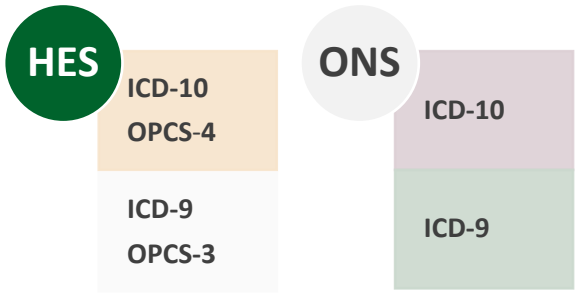


CPRD GOLD coding systems



Coding variable (applicable file)	Description	Code list for mapping
Medcode (all files, except therapy)	CPRD unique code for the readcode or medical term selected by the GP. Provides consistent coding and less errors. E.g. medcode is always numeric rather than mixed upper lower letters etc. as in readcodes	Medical.txt
Prodcode (therapy file)	Any drug products and appliances using the Gemscript product code system. E.g. prescriptions, medications. Add prefix “d_”	product.txt
Bnfcode, qty, numdays, numpacks (therapy file)	Code representing the chapter & section from the British National Formulary for the product selected by GP + different combination of dosage, timings and frequency	BNFCodes
Entity (clinical, test file)	“enttype”. Defines data1 to data8. Contains everything from death/birthdates, and birthweight to scores on self-report measures. Depends on file type.	Entitiy.
• Scale and measures (clinical, test file)	“enttype”. Same as above, but good to note that data1-data8 contains self-report data. So you can ascertain health presentations from self-report measures and apply e.g. cut-off to establish potential diagnosis Example:	Entitiy + scoremethod

HES and ONS



ICD coding of mainly discharge reports completed by professional hospital coders following guidance.

- *Can only code actual diagnoses not suspected events (very important for violence-related presentation)*

ONS

- 1979-2000 uses ICD-9 thereafter ICD-10
- Important! A few publications with corrections, as they missed this.

HES

- Up to 1998 ICD-19 and OPCS-3, and thereafter ICD-10/OPCS-4
- See Herbert et al. (2017), or CLOSER resource more info

Data preparation

Before applying indicators

Chucking data vs. Masterfile (one to rule them all)

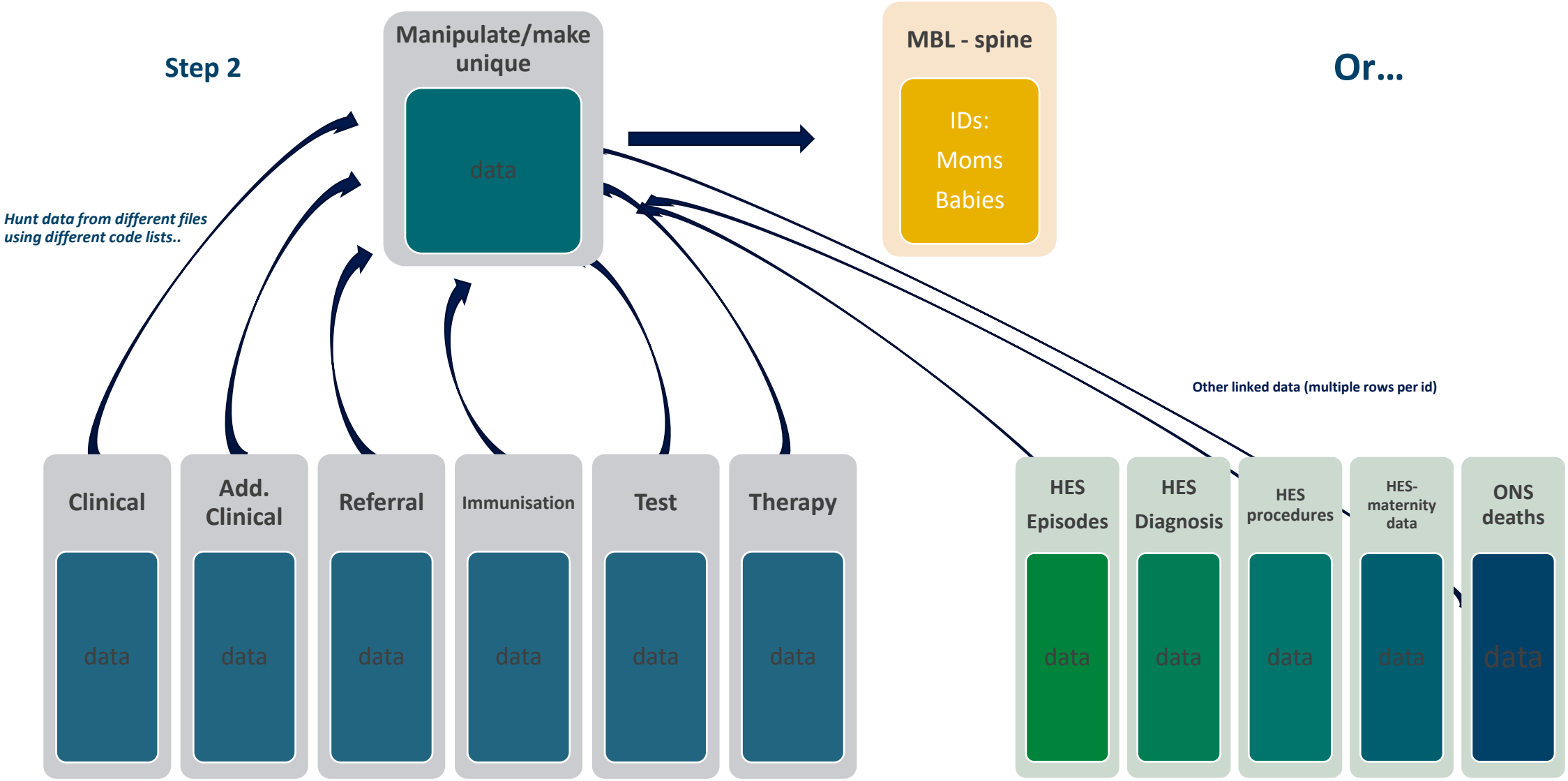
Two main options to obtain ACEs in different data source:

- **Chunking data**: hunter gather chunks from each file individually using specific code lists and merge back to a “spine” or vice versa. This is preferred when you are working with very larger files, and unable to load required files in into working memory (e.g. R).

- **Masterfile**: combine all relevant files into one “master file”. This option is easier when working with smaller or moderate sized files (<10 million patients), as you have all data in one place structured the same way. However, it takes some careful planning in structure and knowing how to handle the file.

- Once you built a master file, you simply gather data from one file rather than multiple. Or for a smaller set of patients you could merge the ACE code list onto the Masterfile directly…

Option 1: chunking data – take what you need from each file



Option 2: Masterfile

Step 2

Bind all data into one file, one to rule them all

Master file (everything in one file)

- Clinical
- Referral
- Immunisation
- Test
- Therapy
- HES episodes
- HES Diagnosis
- HES procedures
- ONS deaths etc

Hunt data from ONE file using ONE code list..

Manipulate/make unique

data

MBL - spine

IDs:
Moms
Babies

Clinical
data

Add. Clinical
data

Referral
data

Immunisation
data

Test
data

Therapy
data

HES Episodes
data

HES Diagnosis
data

HES procedures
data

HES-maternity data
data

ONS deaths
data

Option 2. Creating a master file: example structure

Skip to next slide for option 1 “chunking”

1. **Make sure all file columns follows the same structure (e.g. follow file structure of the “CPRD clinical file” which allows for generic data columns).**

2. **Add an extra column to each file to depict original/file source (HES episodes, CPRD clinical)**

3. **Keep only columns you need to save size (e.g. constype, sysdate, data8 can easily be omitted, as they are rarely used).**

Always keep the same				Generic data fields containing different data depending on file							File source
patid	eventdate	medcode	enttype	data1	data2	data3	data4	dataS	data6	data7	source
4578	2018-05-09	12901	3	0	1	s		0	1		c
4578	2018-05-09	A101									he

Both options:
Data preparation before applying code lists

4. Clean and remove any punctuation from codes

5. Make sure all files are in long format (e.g. ONS is supplied in wide format and needs restructuring)

6. Make sure you convert all data to the same class e.g. character, date (in the right order, year-month-date)

7. Make sure to make codes from different coding system unique to avoid deduplication and preserve their meaning (i.e. different coding systems may use the same code for different purposes)

Preserve codes meaning/avoid deduplication

7.1 Prodcodes (i.e. medications/prescriptions) and ICD-9 codes have thousands of codes that are exactly the same as the medcodes (i.e. diagnoses/symptoms) but they mean different things.

For example:

- CPRD prodcode: 11246 – "Lofexidine 200microgram tablets" =
- CPRD medcode: 11246 – "At risk violence in the home"

7.2 In order to bind different coding systems into a master file, we first need to make each coding unique so we can preserve their meaning when we run a code list later on

7.3 For prodcodes, we can add the prefix "d_" to the coding column to the therapy file and your code list. Similarly, for ICD-9 add the prefix, "e_".

For example:

- prodcode: 11246 = **d**_11246

Now unique from medcode

Ascertaining ACEs

How to

Step 1 - Extract a smaller “study specific file” from the relevant data sources

Whilst using a Masterfile option provides everything in one place, it is often too big and too slow to work with directly in R. So you may still need to “chunk” this file down a little. The benefits of using a masterfile still applies, as you have everything in one place for any needed data manipulation.

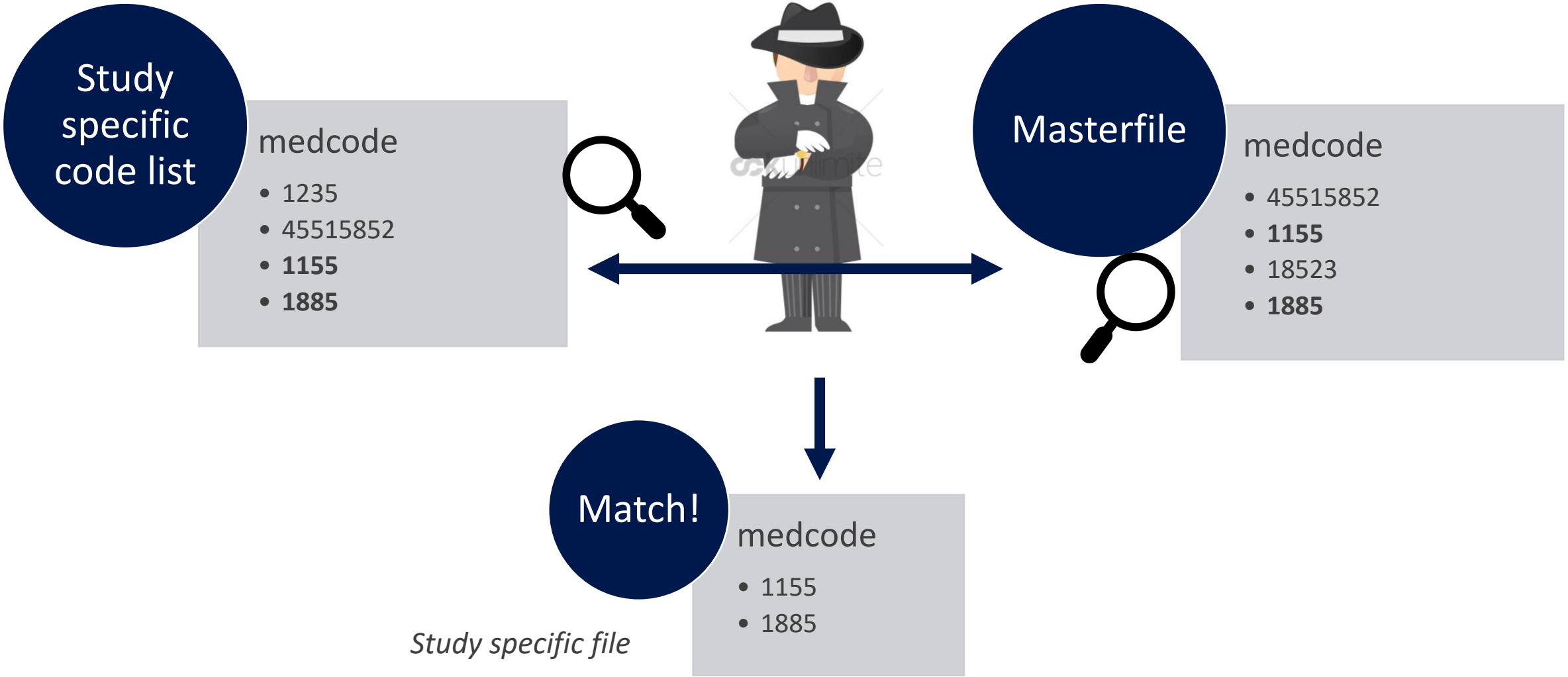
1. **Create a smaller study specific file containing only relevant data which R can easily work with:**

    - Crossmatch your study specific code list and your patient IDs against the Masterfile using e.g. “fastMatch package”.

    e.g. variable1 %fin% datafile1$variable & variable2 %fin% datafile2$variable etc..

    - The retained data frame/file becomes your study specific file which R can work with using a “hunter, gather, merge cycle”.

fastMatch package *(%fin%) example* - Effectively extract relevant data to your code list



Example using dplyr and fastmatch package: *match multiple vectors/variables (codes + IDs) at the same time*

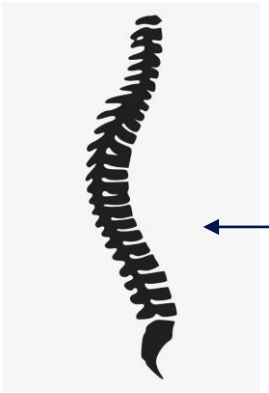
- matches_retained_data **<-** masterfile **%>% filter**(medcode **%fin%** inurycodes**$**medcode **&** patid **%fin%** mbl**$**babypatid)

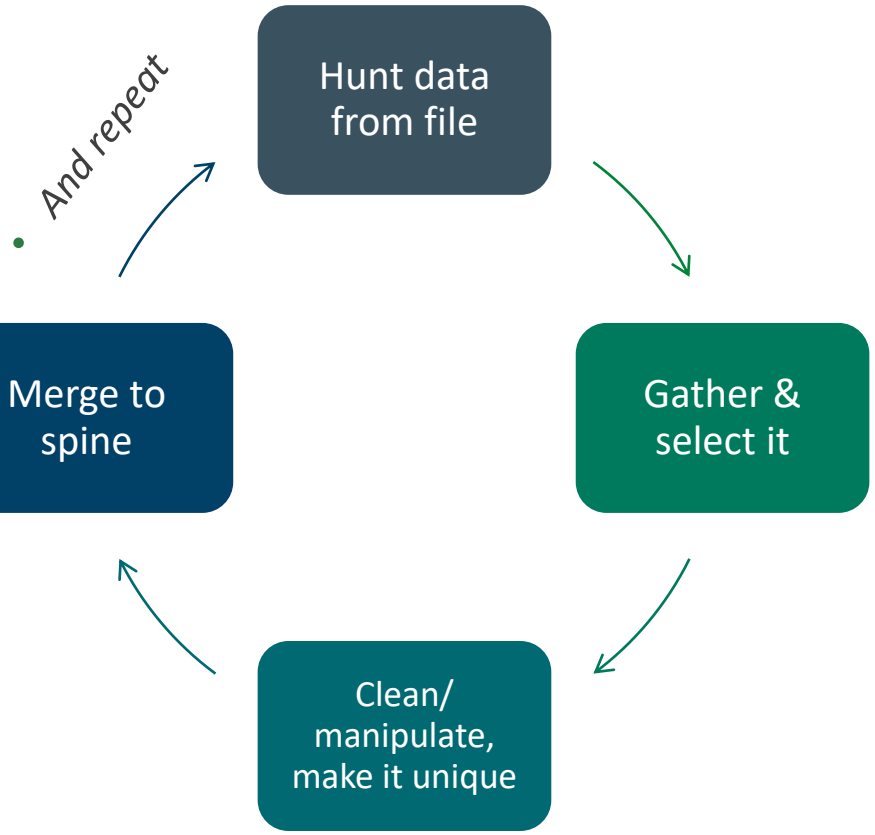
- **English translation:** Take the master file **AND then RETAIN** medcodes **MATCHing** medcodes in my study specific codelist (e.g. injuries) **AND keep only those matches where** patient Ids **MATCHes** patient ids in the mother babylink. **ASSIGN/save** all of this stuff to a new data frame called "matches_retained_data"

- **$**= tells R which column/variable from another dataframe outside of the current pipeline it should sue
- *$ not needed for any columns/variables already in the pipeline by dplyr i.e. selected dataframe. In the above, for example, we say take Masterfile AND then (%>%)…so the Masterfile is already in the pipeline and no need to tell R to look elsewhere for a variable.*

Step 2. Hunter, gather and merge cycle

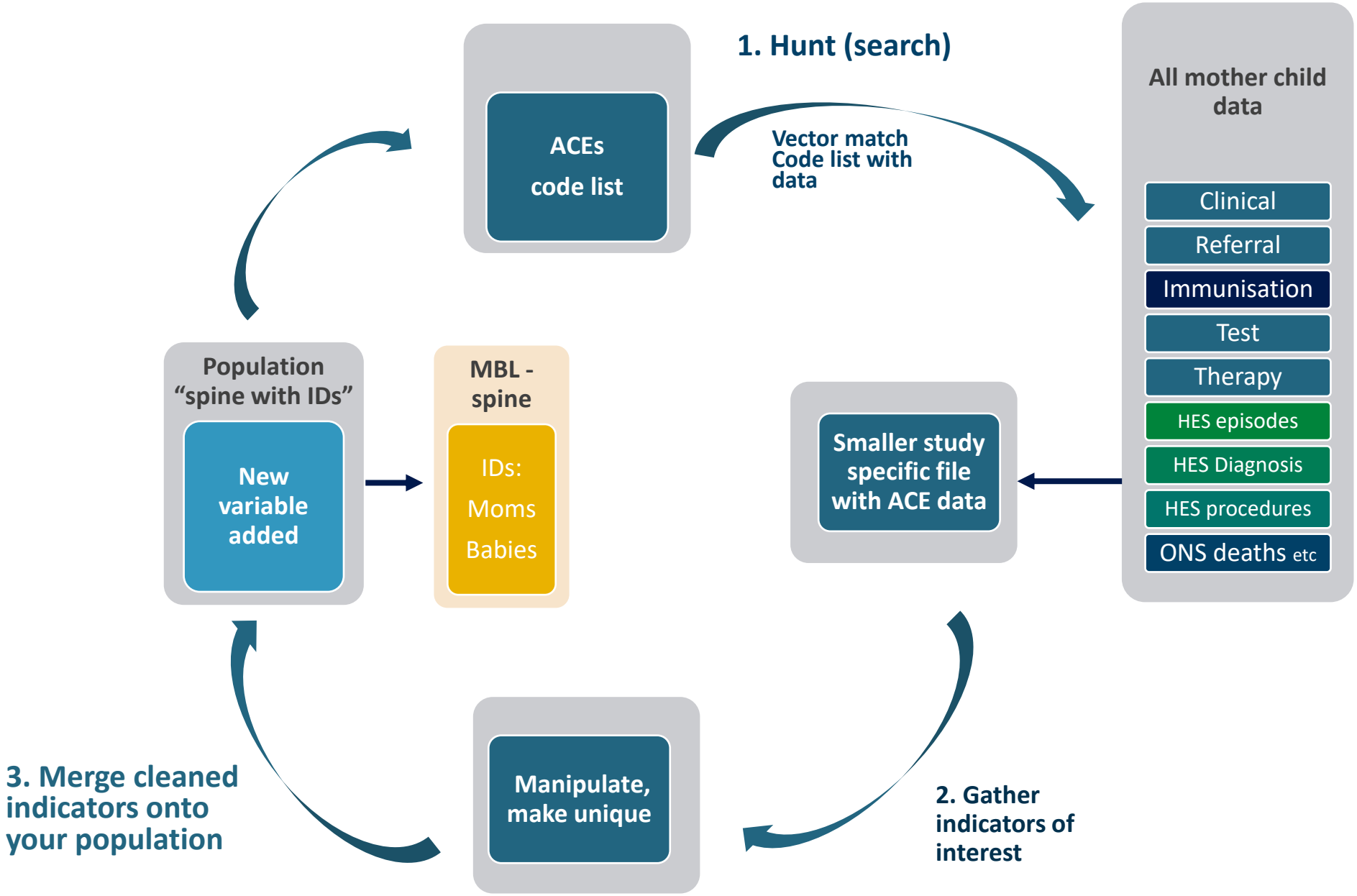
- *And repeat*



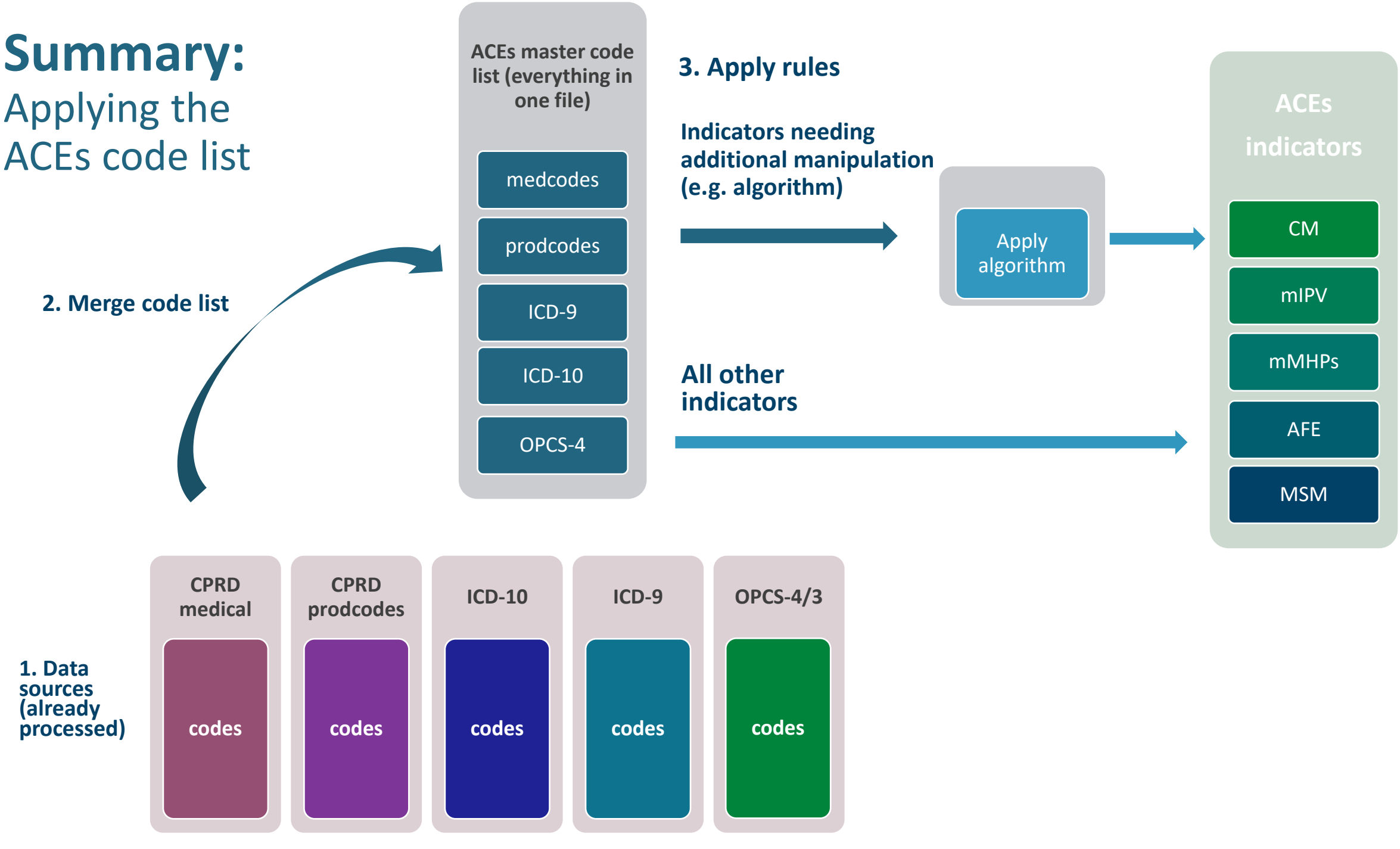


Step 2 - Hunter, gather and merge cycle

- *And repeat*



Summary:
Applying the ACEs code list



R crib sheet for tutorial

Tidyverse, datatable and fastmatch

Command	Description	Examples	Package
fread()	Loading comma or tab delimited data with specific criteria e.g. lazy loading (only 100 rows)	mbl <- fread("MBL_DM.csv",colClasses="character",header=T) mbl <- fread("MBL_DM.csv",colClasses="character",header=T, nrows=100)	data.table
fwrite()	Saving data into comma or tab delimited data	fwrite(mbl,"MBL_Mum_19_162R_DM_v1.csv")	data.table
%>%	Pipeline operator (AND THEN)	mbl <- mbl %>% rename(patid=mumpatid) %>% mutate_all(as.character)	dplyr
select()	Select specific columns/variables	tomerge_mom <- patient %>% select(patid,marital,frd,crd,toreason)	dplyr
rename()	Rename variable	mbl <- mbl %>% rename(patid=mumpatid)	dplyr/tidyr
filter()	Subset or retain data by specific criteria can be combined with fastmatch	mbl <- mbl %>% filter(age>=0)	dplyr
mutate() / mutate_all()	Create new variable or overwrite	immu <- immu %>% mutate(source="i",enttype="") %>% mutate_all(as.character)	dplyr/tidyr
distinct()	Keep only unique value by specific variable	clinical <- clinical %>% distinct(babypatid,medcode,eventdate,.keep_all = T)	dplyr/tidyr
rbindlist(list())	Bind data vertically, make sure columns are the same	masterfile <- rbindlist(list(clinical,ons,hes_all,ref,test,immu,the))	data.table
left_join()	Fast left merging by specific variables	mbl <- left_join(mbl,tomerge_mom,by="patid",all.x=T,copy=F)	tidyr
join_all()	Fast full merging keeping left/right/courtesan etc.	mbl <- join_all(list(mbl, closesttobirth_injury),by="babypatid",type="full")	plyr
melt()	Convert data from wide to long format	deaths_long <- melt(ons,id.vars=c("patid","dod"))	data.table
dcast()	Convert from long to wide format	deaths_wide <- dcast(deaths_long,patid+dod~variable,value.var = "value")	data.table
gsub()	Replace with pattern/clean/remove	codes$medcode <- gsub('\\s+','',codes$medcode) codes$medcode <- gsub('\\.','',codes$medcode) codes$medcode <- gsub('[^[:alnum:]]','',codes$medcode)	R base package
%fin%	Fast vector/variable/pattern match, use to retain matches e.g. codelist against another vector e.g masterfile$medcode	anyinjury <- masterfile %>% filter(medcode %fin% inurycodes$medcode) child_injury <- masterfile %>% filter(medcode %fin% inurycodes$medcode & patid %fin% mbl$babypatid)	fastmatch
ifelse()	replace value if meeting assigned criteria e.g. missing values with 0	mbl$injury <- ifelse(is.na(mbl$injury), "0“, mbl$injury)	R base package

Please note: R is limited in working memory.

ff package overcomes limitations in working memory by using on disk format () to bind multiple files into a master file.