Supplementary Note 1. Methods

We performed three analysis steps to move from individual level data on Read chapters to a dissimilarity matrix which was used to derive the networks and hierarchical dendrograms presented in this study. Here we describe these steps in more details:

1. Linear mixed-effects logistic regression to estimate association between two Read chapters
2. General linear hypothesis to derive Jaccard distance at specified covariate values
3. Hierarchical cluster analysis and undirected network analysis

Linear mixed-effects logistic regression

We used a generalised linear mixed model (GLMM) to estimate the probability of recording an event in both Read chapters for all combinations of the 19 chapters included in this study (hence 171 regression models). For each regression let $y_{iq} = 1$ when an individual patient i from practice q records events in both Read chapters being modelled. Then the regression for each pair of Read chapters in matrix notation is as follows:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon} \tag{1}$$

Where $\boldsymbol{y}$ is an $N \times 1$ vector of binary outcomes where N is the total number of individuals in the model and $y = 1$ if there are records from both chapters and $y = 0$ if there is only a record from one chapter (Patients with no record from either chapter were removed for each regression). Then $\boldsymbol{X}$ is an $N \times 3$ matrix of the fixed effect predictor variables; exposure status (0 = unexposed, 1 = exposed), age at matching date (centred on age 50), and duration of follow up following the first event (centred on 5 years of follow up). The vector $\boldsymbol{\beta}$ is then then the vector of fixed-effects regression coefficients. $\boldsymbol{Z}$ is the $N \times q$ design matrix for the q random effects corresponding to the unique GP practices contributing patients to the corresponding regression. The vector $\boldsymbol{u}$ then captures the random effect and $\boldsymbol{\epsilon}$ is an $N \times 1$ colum vector of the residuals.

Since patients without a record in either Read chapter under study could not contribute to the likelihood, the size of each regression was different, with varying numbers of individual patients N and practices q. We used the `glmer` function from the `lme4` package in R:

```
mf <- paste("y ~ ca + age50 + exy5 + mu")
m1r <- glmer(as.formula(paste(mf,"+ (1|pr)")), data=DD, nAGQ=0, family=binomial())
```

Where `y` is the outcome for patient i in practice q, and `pr` is the practice identifier for q. The fixed effects are estimated for the variables *ca*, a binary indicator for exposure status; `age50`, age at study entry centred on 50; `exy5`, follow up after the first event centred on 5 years; and `mu`, sex of the participant. We ran the model using a logit link function and set the integer scale (`nAGQ`) to zero. This improved performance of the model and reduced computational burden but provides a less exact form of parameter estimation for GLMMs by optimising the random effects and the fixed-effects coefficients in the penalized iteratively reweighted least squares step.

General linear hypothesis

To convert our regression output into probabilities we used a general linear hypothesis at relevant parameters values for the fixed effects in our model. We assumed that age and sex would play an important role in the probability of multimorbidity, as well as exposure status. Therefore we derived separate probabilities at pre-specified ages and for both sexes included in this study, and both exposure status (giving $2 \times 2 \times 2 = 8$ networks for eczema and asthma).

To extract the probability of a record from both Read chapters in teh model, we used the `glht` function from the `multcomp` package in `R`. The linear hypothesis we tested for each of the eight evaluations was as follows:

```
K <- c("(Intercept) = 0",
       "(Intercept) + caTRUE = 0",
       "(Intercept) + muTRUE = 0",
       "(Intercept) + muTRUE + caTRUE = 0",
       "(Intercept) - 32*age50 = 0",
       "(Intercept) - 32*age50 + caTRUE = 0",
       "(Intercept) + muTRUE - 32*age50 = 0",
       "(Intercept) + muTRUE - 32*age50 + caTRUE = 0")
```

This evaluated the probability of records in both Read chapters in the model for the following respecively:

- Unexposed men at age 50 `(Intercept) = 0`
- Exposed men at age 50 `(Intercept) + caTRUE = 0`
- Unexposed women at age 50 `(Intercept) + muTRUE = 0`
- Exposed women at age 50 `(Intercept) + muTRUE + caTRUE = 0`
- Unexposed men at age 18 `(Intercept) - 32*age50 = 0`
- Exposed men at age 18 `(Intercept) - 32*age50 + caTRUE = 0`
- Unexposed women at age 18 `(Intercept) + muTRUE - 32*age50 = 0`
- Exposed women at age 18 `(Intercept) + muTRUE - 32*age50 + caTRUE = 0`

This function will return the prediction on a logit scale for the probability p of a record from both Read chapters in the model at the corresponding fixed effects values. To convert this estimate (call it $x = log\left(\frac{p}{1-p}\right)$) to a probability we take:

$$p = \exp\left(\frac{x}{1+x}\right)$$

These steps results in an estimated probability for a record from two Read chapters for each of the 171 combinations of chapters and the 8 estimated linear hypothesis. We then converted this into a dissimilarity matrix for the hierarchical cluster analysis.

One important note, since the variable `exy5` is not included in these linear hypotheses, the fixed effect of follow up was controlled for at a centred value of 5 years. In other words, these hypotheses test for the probability that a man/woman, exposed/unexposed, 18/50 year-old, would record a Read code in the other chapter within 5 years of a record in the first (for each pair of studied Read chapters).

Hierarchical cluster analysis and undirected network analysis

We constructed a dissimilarity matrix (8×190) for the probability of multiple records from each combination of two Read cahpters. There were 8 for combinations of the two specified ages (18 and 50 years-old), sexes (male and female) and exposure statuses (exposed and unexposed). There were 190 rows in the matrix for the 171 possible discordant combinations plus the 19 chapters themselves.

This dissimialirty matrix was then analysed using the `hclust` function from the `stats` package using the complete linkage clustering method. This performs a hierarchical cluster analysis on a set of dissimilarities for the 19 Read chapters being clustered at each of the 8 combinations of explanatory variable values.

In short, each object is assigned to its own cluster and then the algorithm proceeds iteratively, at each stage joining the two most similar clusters, continuing until there is just a single cluster. At each step the two clusters seperated by the "shortest distance" are combined, in our example this is the shortest Jaccard distance (or highest probability of a record from both Read chapters in pairs). The method is an agglomerative scheme that erases rows and columns in the matrix as old clusters are merged into new ones.

The `hclust` generates a hierarchical tree produced by the clustering process. This tree is then used to plot the dendrograms in this study.

The undirected network graphs were constructed using the `graph_from_data_frame` function from the `igraph` package where the node size is determined by the number of recorded events in each specific Read

chapter for each combination of age, sex and exposure in our analysis. The edges are determined by the conditional probability of a record in one chapter given a record in another, from the GLMM analysis above.