# A systematic analysis of the contribution of genetics to multimorbidity and comparisons with primary care data

Olivia Murrin* [1], Ninon Mounier* [1], Bethany Voller* [1], Linus Tata* [2], Carlos Gallego-Moll [3,4], Albert Roso-Llorach [3,4], Lucía A Carrasco-Ribelles [3,4,5], Chris Fox [6], Louise M Allan [6], Ruby M Woodward [7], Xiaoran Liang [1], Jose M Valderas [6,8], Sara M Khalid [9], Frank Dudbridge [7], Sally E Lamb [6], Mary Mancini [10], Leon Farmer [10], Kate Boddy [6], Jack Bowden [1], David Melzer [1], Timothy M Frayling # [1,11], Jane AH Masoli # [1,12], Luke C Pilling # [1], Concepción Violán # [4,5,13,14,15], João Delgado # [1]

* = joint first authors

# = joint senior authors

## Affiliations

1. Department of Clinical and Biomedical Sciences, Faculty of Health and Life Sciences, University of Exeter, UK
2. Research Software Engineering Group, University of Exeter, UK
3. Unitat de Suport a la Recerca Metropolitana Nord, Fundacio Institut Universitari per a la recerca a l'Atencio Primaria de Salut Jordi Gol I Gurina (IDIAPJGol), Barcelona, 08007, Spain
4. Grup de REcerca en Impacte de les Malalties Cròniques i les seves Trajectòries (GRIMTra) (2021 SGR 01537), Institut Universitari d'Investigació en Atenció Primaria Jordi Gol (IDIAPJGol), Mare de Déu de Guadalupe, 2, Barcelona, 08303, Spain
5. Red de Investigación en Cronicidad, Atención Primaria y Prevención y Promoción de la Salut (RICAPPS), Instituto de Salud Carlos III (ISCIII), Avenida Monforte de Lemos,5, Madrid, 28029, Spain
6. Department of Health and Community Sciences, Faculty of Health and Life Sciences, University of Exeter, UK
7. Department of Population Health Sciences, University of Leicester, UK
8. Department of Family Medicine, National University Health System, 1E Kent Ridge Road, Singapore 119228
9. Centre for Statistics in Medicine, Nuffield of Orthopaedics Rheumatology and Musculoskeletal Sciences, University of Oxford, UK
10. Public and Patient involvement representative
11. Department of Genetic Medicine and Development, Faculty of Medicine, 1 rue Michel-Servet, CH-1211 Genève 4, Switzerland
12. Royal Devon University Healthcare NHS Foundation Trust, Barrack Road, Exeter, EX2 5DW, UK
13. Unitat de Suport a la Recerca Metropolitana Nord, Institut Universitari d'Investigació en Atenció Primaria Jordi Gol (IDIAP Jordi Gol), Mare de Déu de Guadalupe, 2, Mataró 08303, State, Spain
14. Germans Trias i Pujol Research Institute (IGTP), Street, Badalona, 08916, State, Spain
15. Department of Medicine, Universitat Autònoma de Barcelona, Plaça Cívica, 1, Cerdanyola de Vallès, 08193, State, Spain

## Corresponding authors

Prof Timothy Frayling (Timothy.Frayling@unige.ch)
Dr João Delgado (J.Correa-Delgado@exeter.ac.uk)

## Abstract

### Background

Multimorbidity, the presence of two or more conditions in one person, is increasingly prevalent. Yet shared biological mechanisms of specific pairs of conditions often remain poorly understood. We address this gap by integrating large-scale primary care and genetic data to elucidate potential causes of multimorbidity.

### Methods

We defined chronic, common, and heritable conditions in individuals aged ≥65 years, using two large representative healthcare databases [CPRD (UK) N=2,425,014 and SIDIAP (Spain) N=1,053,640], and estimated heritability using the same definitions in UK Biobank (N=451,197). We used logistic regression models to estimate the co-occurrence of pairs of conditions in the primary care data. Linkage disequilibrium score regression was used to estimate genetic similarity between pairs of conditions. Meta-analyses were conducted across healthcare databases, and up to three sources of genetic data, for each condition pair. We classified pairs of conditions as across or within-domain based on the international classification of disease.

### Findings

We identified N=72 chronic conditions, with 43·6% of 2546 pairs showing higher co-occurrence than expected and evidence of shared genetics. Notably, across-domain pairs like iron deficiency anaemia and peripheral arterial disease exhibited substantial shared genetics (genetic correlation $R_g$=0·45[95% Confidence Intervals 0·27:0·64]). N=33 pairs displayed negative genetic correlations, such as skin cancer and rheumatoid arthritis ($R_g$=-0·14[-0·21:-0·06]), indicating potential protective mechanisms. Discordance between genetic and primary care data was also observed, e.g., abdominal aortic aneurysm and bladder cancer co-occurred but were not genetically correlated (Odds-Ratio=2·23[2·09:2·37], $R_g$=0·04[-0·20:0·28]) and schizophrenia and fibromyalgia were less likely to co-occur but were positively genetically correlated (OR=0·84[0·75:0·94], $R_g$=0·20[0·11:0·29]).

### Interpretation

Most pairs of chronic conditions show evidence of shared genetics and co-occurrence in primary care, suggesting shared mechanisms. The identified shared mechanisms, negative correlations and discordance between genetic and observational data provide a foundation for future research on prevention and treatment of multimorbidity.

### Funding

UK Medical Research Council [MR/W014548/1].

## Introduction

Multimorbidity is the coexistence of two or more long-term conditions (LTCs) and is a growing problem globally, projected to affect two-thirds of 65-year-olds in the next decade. [1] Multimorbidity has been proposed as a clinical marker of accelerating ageing, resulting in increased frailty and mortality. Clinical management of multiple co-existing LTCs places a disproportionate economic and capacity burden on healthcare systems, particularly in the context of ageing populations. [2–6] Clinical research, including trials focusses on single conditions, and research to understand, prevent and treat multimorbidity is needed. The mechanisms driving the accumulation of LTCs and resulting in multimorbidity remain uncertain.

The majority of previous multimorbidity research is observational, focusing on the concurrence of conditions as a raw or weighted count of selected conditions. Disease clustering approaches have revealed common combinations of LTCs, such as cardiovascular and metabolic, with variability in the conditions chosen and therefore the identified clusters. [7] Clustering-based approaches favour high-prevalence LTCs, such as hypertension, while less prevalent LTCs remain understudied. [8,9] Alternative approaches have analysed disease pairs, using genetic correlations to identify potential shared genetics. However, most genetic studies have focused either on small sets of closely related conditions, [10] single large datasets such as the UK Biobank, or genetic correlations across multiple traits using broad disease definitions. [11]

Genetic data provide the opportunity to understand underlying mechanisms with less susceptibility to confounding, bias and reverse causality than observational studies. Summary-level genetic association data is available for many common conditions based on studies including 10,000s to 100,000s cases. This genetic data enables the comparison of conditions across, as well as within, patients and datasets. For example, linkage disequilibrium (LD) score regression is an approach that can be used to compare genetic similarity between conditions from separate case-control studies of separate diseases, even when each study has no information on the other condition. [11,12] A systematic, multi-modal data investigation remains necessary to identify potential biological mechanisms and opportunities for the prevention and treatment of multimorbidity.

In this study, we combine the advantages of healthcare datasets (representativeness and very large sample size) with the advantages of genetic studies (very large sample size, less confounding and reverse causality) to further understanding of multimorbidity. In contrast to most previous studies, we analyse specific pairs of LTCs, rather than counts or clusters of conditions, and meta-analyse data from more than one source. We provide a comprehensive assessment of the genetic and observational relationships between 72 LTCs common in people aged over 65 years - an analysis comprising 2,546 pairs. Our study identifies hundreds of pairs of conditions with shared genetic factors, many from across traditional disease domains; compares genetic evidence of the co-occurrence of conditions with observational evidence from healthcare data; and provides a powerful resource for the study of multiple long-term conditions in multiple datasets.

## Methods

### Design and populations

### Primary care healthcare data

Observational data were obtained from two independent databases of electronic health records: 1) The UK Clinical Practice Research Datalink (CPRD), including data from >30 million National Health Service (NHS) patients from >700 primary care practices.[13] 2) The Spanish Information System for Research in Primary Care (SIDIAP), including data from 6 million patients from 328 practices.[14] Clinical records in both the UK and Spain undergo standardised coding to contribute to disease registers and healthcare planning. In SIDIAP the electronic codes are recorded using the International Classification of Disease, version 10 (ICD-10). In CPRD codes are a mixture of Read v2 Codes, EMIS and SNOMED codes.

### Genetic

Genetic data from 3 sources were included 1) UK Biobank (UKB),[15] a large cohort study with 500,000 individuals with baseline data linked to electronic health records; we used 450,197 individuals genetically similar to the 1000 Genomes European reference population ("EUR-like" - see Supplementary Information). 2) FinnGen, a large-scale genomics initiative linking diagnosis to genotype data in 377,277 participants (release 9).[16] 3) Published disease-specific Genome Wide Association Study (GWAS) meta-analyses summary statistics. For these disease-specific GWAS, we used the European Bioinformatics Institute (EBI) GWAS Catalog,[17] disease-specific public repositories, and, if necessary, contacted authors of GWAS to obtain summary association statistics.

### Defining Long-Term Conditions

LTCs were defined and selected for analysis based on chronicity, prevalence, and heritability as detailed below. Selected conditions were assigned to disease domains based on the ICD-10 chapters: https://icd.who.int/browse10/2019/en.

***Step 1: Selection of conditions – chronicity***

We included only LTC or with sequelae lasting more than 3 months to meet the definition of multimorbidity. The definition of disease chronicity was adapted from chronic conditions published by Calderón-Larrañaga et al.[18] The LTC code lists were compared with CALIBER interoperable code lists,[19] and adapted with clinician input to refine and remove acute diagnostic codes. This process generated 232 LTC code lists (Supplementary Figure 1).

***Step 2: Selection of conditions – prevalence***

Prevalence was estimated using CPRD and SIDIAP data for individuals aged 65 and older who were alive and registered on the 1st of January 2020. These databases are population-representative sources of primary care data, providing the most comprehensive set of long-term conditions, given many are not treated in hospital. The cutoff date minimises the impact of the COVID-19 pandemic in our analyses. We included conditions with a prevalence greater than 0.5%, supplemented by clinician and Patient and Public Involvement and Engagement (PPIE), resulting in 84 LTCs out of 232 proceeding to step 3 (see PPIE section below).

***Step 3: Selection of conditions - Heritability***

We performed GWAS for 84 LTCs in UKB, using our defined clinical code lists and the REGENIE software (v3.1.3).[20] See the Supplementary Information for details. Briefly, analyses were adjusted for age, sex, genotyping chip, and assessment centre. We restricted genetic variants to those with a

minor allele frequency (MAF) of >0·1%, and an imputation INFO score ≥0·3. SNP-based heritability was estimated using GWAS summary statistics and LD-score regression (LDSC).[12] We used the 1000 Genomes EUR reference population LD data throughout. 72 LTCs met criteria for analysis (Supplementary Figure 1).

**Co-occurrence of LTCs in primary care**

Logistic regression models tested the likelihood of two LTCs co-occurring in observational data, with models adjusted for age and gender, a Benjamini-Hochberg correction was used to account for multiple testing (additional detail in supplemental information). Associations (i.e. odds-ratios) between LTCs were estimated in CPRD AURUM and SIDIAP, and meta-analysed with fixed-effects using the RMA function in the R package 'metafor'.[21]

**Genetic Correlation**

To provide the most powerful set of genetic data we sought to meta-analyse genetic studies according to two criteria: first, there should be evidence that the conditions defined in the different genetic studies were the same, or very similar, conditions. Second, there should be no overlap between the sources of genetic data. We used LDSC to estimate the within-condition between-dataset genetic correlation ($R_g$) and limited meta-analyses to studies where the within-condition genetic correlation was >0·8.[11] To ensure there was no overlap in genetic data, The FinnGen and Consortium data were added to the meta-analysis when within-condition $R_g$ with UKB was >0·8, otherwise we used UKB only (i.e., the genetic evidence suggests that the conditions used by consortium GWAS and/or FinnGen are not consistent with those defined using diagnostic codes for LTCs in our project for UK Biobank). Where Consortium data included UKB and/or FinnGen, Consortium data were favoured, because of the larger number of cases, and UKB/FinnGen excluded to avoid sample overlap. Studies were meta-analysed using GWAMA.[22] LTC pairs with genetic correlation at a false discovery rate of 5% were first separated into those within and across domains and second, sorted into three groups as those with weak, intermediate, and strong genetic correlations. See the Supplementary Methods for individual study references, Supplementary Figure 2 for analysis flowchart, and Supplementary Table 1 for effective sample size and other information.

**Comparison between observational and genetic correlations**

Linear regression models estimated the association between genetic correlations and observed co-occurrence between LTC pairs, where LTC pair genetic correlation was used as the dependent variable and observed co-occurrence as the outcome. Additional models estimated the associations in subsets of LTC pairs within and across disease domains. LTC pairs were classified as within or across disease domains based on ICD-10 chapters (Supplementary Table 2).

**Professional, Patient and Public Involvement**

Two co-authors (LF and MM) are public collaborators with direct experience of living with multiple LTCs. They are co-investigators attending fortnightly research meetings to co-develop the research. Additional workshops for patients and carers with experience of multiple LTCs have contextualised the importance of the research and directly informed research decisions on LTC selection and refinement as outlined above.[23]

Healthcare professionals, including primary and secondary care physicians and allied healthcare professionals (led by authors JM, CVF and SEL) informed the definition, selection, and precision coding of LTCs. Detailed clinician review of LTC pairs identified potential underlying mechanisms and selected LTC pairs where an association was considered novel.

**Ethical considerations**

This study was approved by the relevant ethics committees:

SIDIAP Scientific and Ethical Committees (19/518-P) on 18/12/2019. The SIDIAP database is based on opt-out presumed consent. If a patient decides to opt out, their routine data would be excluded of the database.

CPRD ISAC committee protocol number 23_003109.

The Northwest Multi-Centre Research Ethics Committee approved the collection and use of UK Biobank data for health-related research (Research Ethics Committee reference 11/NW/0382). UKB was granted under Application Number 9072.

**Role of the funding source**

The funders had no input into the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the paper for publication.

## Results

### The majority of long-term conditions common in older people are heritable

We identified 72 LTCs that were common and showed evidence of heritability. The most prevalent of these conditions in primary care data were hypertension (CPRD: 51·7%; SIDIAP: 63·2%), osteoarthritis (CPRD: 35·3%; SIDIAP: 37·8%), and upper body enthesopathy (CPRD: 27·0%; SIDIAP: 20·0%) (Supplementary Table 1). The most heritable of these LTCs were fibromyalgia ($h^2$=25·1%) and type 2 diabetes ($h^2$=21·3%) (Supplementary Table 1). Further details of all 72 plus 12 additional conditions that did not meet our heritability criteria are available online: https://gemini-multimorbidity.shinyapps.io/atlas.

### Associations between conditions

Pairwise combinations of 72 conditions resulted in 2,546 pairs, of which 260 were within-domain and 2,286 across-domain (based on ICD10 classification) (Figure 1; Table 1). Ten pairs were excluded due to overlapping code lists, e.g., transient ischaemic attack (TIA) and "all stroke". Prevalence estimates were based on data from 2,425,014 CPRD patients (46·3% male) and 1,053,640 SIDIAP patients (43·0% male). Genetic correlations were estimated using data from 450,197 UKB participants (45·7% male), 377,277 FinnGen participants (44·1% male), and from condition-specific GWAS data with sample sizes ranging from 12,366 to 181,522 cases.

Regression analyses across N=2,546 pairs demonstrated that a 1% increase in $R_g$ equates to a 2·76% increase in the odds of co-occurrence (95% Confidence Intervals 2·65%-2·87%), thus the stronger the genetic correlation between an LTC pair, the higher the chance of the pair co-occurring in primary care.

#### *1. The majority of within-domain LTC pairs are genetically correlated and co-occur together in primary care data*

From 260 pairs of within-domain LTCs, N=209 (80·4%) were genetically correlated and co-occurred in primary care more often than expected (e.g., heart failure and atrial fibrillation - $R_g$=0·60-SE=0·03, OR: 7·53 [7·45-7·61]). Significant negative genetic correlations were absent in this group (Figure 2, blue).

#### *2. Many across-domain LTC pairs are genetically correlated and tend to co-occur in primary care data*

In across-domain pairs, we identified N=105 (4·6%) that were as strongly genetically correlated ($R_g$>0·48) as the strongest third of within-domain pairs. These pairs were more likely to co-occur in primary care (Figure 2, green). For example, sinusitis and gastro-oesophageal reflux disease (GORD) ($R_g$=0·49-SE=0·06), erectile dysfunction and peripheral neuropathy ($R_g$=0·49-SE=0·10) and iron deficiency anaemia and peripheral arterial disease ($R_g$=0·45-SE=0·10) were as strongly associated as within-domain pairs coronary heart disease and stroke ($R_g$=0·49-SE=0·03), rheumatoid arthritis and polymyalgia rheumatica ($R_g$=0·48-SE=0·09), and TIA and peripheral arterial disease ($R_g$=0·48-SE=0·08). Diseases of the musculoskeletal system or connective tissue were present in 8 of the 10 across-domain LTC pairs with the strongest genetic correlations, but only 18.3% of all significant across-domain pairs.

Clinician reviews of LTC pairs highlighted that associations are known, and have established explanatory mechanisms for all within-domain LTC pairs, and for most across-domain LTC pairs. For example, associations could reflect 1) shared pathology (e.g. Type 2 Diabetes and Erectile

dysfunction – OR: 3·27 [3·24-3·30], $R_g$=0·41-SE=0·04), 2) the first condition is a risk factor for the second (e.g. obesity and osteoarthritis – OR: 2·00 [1·99-2·01], $R_g$=0·54-SE=0·02), 3) treatment effects (e.g. intervertebral disc herniation where non-steroidal anti-inflammatory drugs (NSAIDs) could predispose to gastro-oesophageal reflux disease (GORD) -OR: 1·60 [1·58-1·62], $R_g$=0·50-SE=0·03) and 4) shared symptoms leading to overlapping diagnoses (e.g., sinusitis and GORD - OR: 2.00 [1.98-2.02], $R_g$=0·49-SE=0·06). However, other across-domain LTC pairs had no well-established shared mechanisms for genetic and observed associations. Some examples were tendon disorders and diverticular disease (OR: 1·49 [1·48-1·50], $R_g$=0·65-SE=0·13), fibromyalgia and irritable bowel syndrome (IBS) (OR: 3·38 [3·30-3·47], $R_g$=0·65-SE=0·13), fibromyalgia and asthma (OR: 1·87 [1·83-1·92], $R_g$=0·45-SE=0·06) and IBS and peripheral neuropathy (OR: 1·6 [1·57-1·64],$R_g$=0·57-SE=0·12). A total of N=89 genetically correlated and co-occurring LTC pairs involved treatable deficiencies in iron and vitamin B12, e.g. B12 deficiency and COPD (OR: 1.57 [1·55-1·60], $R_g$=0.35-SE=0·06).

***3. A small number of pairs of conditions are negatively genetically correlated and co-occur less often than expected in primary care.***

We identified 33 (1·3% of N=2,546) pairs of LTC conditions that were negatively genetically correlated (FDR <0·05), implying that the genetic risk of one is associated with protection from the other. N=19 (56·0%) of these pairs were observed together less often than expected in primary care. Of the 19 pairs, 16 involved malignancies, for example, skin cancer and rheumatoid arthritis ($R_g$=-0·14-SE=0·04) and five pairs included schizophrenia (e.g. schizophrenia and upper body enthesopathy, $R_g$=-0·17-SE=0·03) (Figure 2, red).

***4. Some pairs of conditions show discordance between observed co-occurrence and genetic correlation***

A total of 34 (1.3% of 2,546) pairs did not co-occur in primary care data more than chance (after false discovery rate correction) but had a positive genetic correlation (Figure 2, purple), such as female genital prolapse and type 2 diabetes (OR: 0·97 [0·96-0·98], $R_g$=0·13-SE=0·02). This group included several pairs involving mental health conditions and pain-related conditions e.g. schizophrenia and fibromyalgia (OR: 0·84 [0·75-0·94], $R_g$=0·13-SE=0·02).

We identified 739 disease pairs (29·0%) showing significant positive co-occurrence without genetic correlation (Figure 2, purple). Examples include fibromyalgia and polymyalgia rheumatica, indicative of explorative diagnosis along a diagnostic pathway (OR: 0·84 [0·75-0·94], $R_g$=0·15-SE=0·12); iron deficiency anaemia with colorectal cancer (OR: 2·55[2·49-2·60] , $R_g$= 0·04-SE=0·07), as well as abdominal aortic aneurism (AAA) with bladder cancer (OR: 2·23[2·09-2·37], $R_g$=0·04-SE=0·12), which may involve incidental findings.

## Discussion

### Associations between conditions and mechanisms of multimorbidity

We report the first, systematic investigation of the shared genetics and co-occurrence in primary care of 2546 pairs of long-term conditions (LTC). Our study builds on previous work in multimorbidity by integrating two of the largest most representative sources of primary care data with large scale genetic data for each of 72 conditions. Most pairs (61·0%) tended to co-occur in primary care data and share genetics, with a subset of across-domain pairs showing genetic correlations as strong as many within-domain pairs. The overall positive relationship between observed phenotypic associations and genotypic correlations is consistent with and considerably extends work from a previous smaller-scale study, limited to 17 conditions in UK Biobank.[24] These findings suggest diverse shared pathways and mechanisms drive co-occurrence of LTCs in multimorbidity.

There are several potential mechanistic explanations for our results. These mechanisms could include shared pathophysiology, such as atherosclerosis, as a likely shared risk factor for cardiovascular diseases such as atrial fibrillation and heart failure.[25] Causal mechanisms may also include one LTC acting as a risk factor for a second LTC, such as obesity increasing the risk of osteoarthritis due to increased mechanical stress on weight-bearing joints. Shared genetic and observational mechanisms between LTCs could result from the combined effect of both types of causal pathways, such as increased risk of erectile dysfunction in type 2 diabetes, caused by shared pathophysiology and by the effect of hyperglycaemia on the endothelium.[26] Lastly, concordance between phenotypic and genetic correlations may result from iatrogenic mechanisms. For example, secondary stroke prevention with antiplatelet medications, such as clopidogrel, increases the risk of gastritis, and the use of non-steroidal anti-inflammatory drugs (NSAIDs) for pain in musculoskeletal conditions such as intervertebral disc herniation increases the risk of gastro-oesophageal reflux disease (GORD).[27]

We highlight LTC pairs that co-occur more often than expected by chance and that are genetically correlated but lack strong clinical understanding. For example, tendon disorder and diverticular disease were associated and novel other than one case study.[28] A few LTC pairs include a readily treatable condition such as B12 deficiency or iron deficiency anaemia, highlighting a potential direct path towards intervention.

LTC pairs showing discordance between genetic correlations and co-occurrence in primary care raise interesting questions about clinical service provision. The 34 LTC pairs with evidence of shared genetics but occurring less often than expected by chance were dominated by combinations of musculoskeletal and mental health conditions, such as schizophrenia with fibromyalgia, and with rheumatoid arthritis, or alcohol addiction and spondylolisthesis. These combinations could suggest that diagnoses of severe mental health conditions lead to underdiagnosis of concomitant physical LTCs probably involving diagnostic overshadowing.[29–31] LTC pairs with increased observed co-occurrence but without evidence of genetic correlation could include clinical instances where diagnosis with one LTC leads to increased odds of being diagnosed with a second LTC. For example, iron deficiency anaemia (IDA) and colorectal cancer, where a diagnosis of IDA triggers an investigation for colorectal cancer; or Abdominal Aortic Aneurysm (AAA) and bladder cancer, where an AAA is incidentally detected during an investigation for bladder cancer. Discordant pairs highlight instances where pathways for diagnosis of chronic conditions are inadequate, and these may remain undiagnosed and untreated.

**Strengths and limitations**

This study uses large and representative data from electronic health records (EHRs), with findings replicated across datasets.[13,14] EHRs are inclusive of those with disability, frailty and MLTC who are frequently under-represented in clinical research. Observational data were also complemented with three high-quality genetics data sources. An exhaustive review process has been carried out to compare chronic, heritable diseases across genetic and EHR data. There is no clear international consensus on the assessment of multimorbidity. In this study, we have used a comprehensive method to select common chronic LTCs in older populations, based on existing literature, that has ensured the inclusion of 2546 multimorbidity disease pairs, the largest systematic collection to date. Clinical experts curated LTC definitions and LTC pairs, which are available for future MLTC investigations.

A few limitations should be considered. 1) The genetic analyses were limited to individuals of European genetic ancestry due to a paucity of large-scale genetic data from people of non-European ancestry. 2) Limiting primary care data to individuals at least 65 years may introduce survival bias; negative effect sizes cannot be automatically conferred as protective. For example, where we found reduced co-occurrence between schizophrenia and musculoskeletal condition another study found an increased rate of mortality between the two.[32] 3) Participants may have co-occurring LTCs because of misdiagnosis or codes associated with the diagnostic pathway, for example, the strong co-occurrence observed for polymyalgia rheumatica and fibromyalgia. 4) Finally, differences in statistical power for defining conditions as genetically correlated or co-occurring means it is important to consider 95% confidence intervals when considering specific pairs. However, in this study we have used extremely large datasets, ensuring tight confidence intervals around our estimates.

**How could these results help clinicians?**

Understanding novel associations and associations involving treatable conditions can highlight opportunities for improved detection, and interventions for prevention, delaying onset or treatment of LTC pairs. Genetic correlations provide a starting point for the identification of specific mechanisms of MLTC providing a foundation for research on potential prevention and treatment. This knowledge can lead to novel treatment approaches and drug repurposing across LTC pairs that will inform clinical guidelines to the benefit of patients.[33]

LTC pairs that are genetically correlated without observed co-occurrence could highlight underdiagnosed conditions, some of which may be amenable to screening or education to improve detection. This includes conditions with a potentially high symptom burden in groups in whom there are barriers to clinical presentation or accessing healthcare, such as people with mental health diagnoses, or highlights conditions that are commonly not diagnosed, such as AAA suggesting screening programs extended only to men over 65 as a one-time event have not been appropriately adopted.[34]

Lastly, we highlight potentially treatable conditions with high co-occurrence, such as iron deficiency anaemia with peripheral neuropathy, and B12 deficiency with diabetes complications and treatments or supplementation advice could be explored further.[35,36]

**Further work**

Future detailed work is planned to investigate the specific relationships between LTC pairs. Multimorbidity represents complex interactions of biological pathways and environmental factors.

Longitudinal research, adjusted for confounders, can elucidate mechanisms and risk factors involved for selected, under-researched LTC pairs with strong genetic correlation. Genetic causal inference methodologies can identify targets for intervention,[37] allowing researchers to test and propose personalised preventative and therapeutic actions.

## Conclusion

We have performed a systematic analysis of multimorbidity, integrating large scale primary care and genetic data from multiple sources, and involving patients as collaborators. We have identified novel combinations of conditions, including those that tend to share genetic factors but not co-occur in primary care, and vice versa. Our data is accessible through an interactive web app (https://gemini-multimorbidity.shinyapps.io/atlas), which we anticipate will provide a valuable resource for further research in multimorbidity.

## Funding

This work was supported by the UK Medical Research Council [grant number MR/W014548/1]. This study was supported by the National Institute for Health and Care Research (NIHR) Exeter Biomedical Research Centre (BRC), the NIHR Leicester BRC, the NIHR Oxford BRC, the NIHR Peninsula Applied Research Collaboration, and the NIHR HealthTech Research Centre. KB is partly funded by the NIHR Applied Research Collaboration South-West Peninsula. JM is funded by an NIHR Advanced Fellowship (NIHR302270). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. CV acknowledges research funding by a "Contratos para la intensificación de la actividad investigadora en el Sistema Nacional de Salud" contract (INT23/00040) from the Spanish Ministry of Science and Innovation.

## Acknowledgements

Data from Clinical Practice Research Datalink (CPRD) was obtained under licence from the UK Medicines and Healthcare products Regulatory Agency. The data is provided by patients and collected by the NHS as part of their care and support. The interpretation and conclusions contained in this study are those of the author/s alone. This research has been conducted using the UK Biobank Resource under Application Number 9072. We want to acknowledge the participants and investigators of the FinnGen study. We also thank the authors and study participants for the published GWAS consortium meta-analyses utilized in this report: see the Supplementary Information for full citations. The authors would like to acknowledge the use of the University of Exeter High-Performance Computing (HPC) facility in carrying out this work.

## Declaration of interests

ARL is now an employee of AstraZeneca and has interests in the company. The work undertaken here was prior to his appointment. SK's group has received funding support from Amgen BioPharma outside of this work. JB is a part time employee of Novo Nordisk Research Centre Oxford, limited, unrelated to this work. TF has consulted for several pharmaceutical companies. All other authors have no disclosures to declare.

## Author contributions

Conceptualization: OM, NM, JMV, FD, SEL, JB, DM, JAHM, LCP, JD. Data Curation: OM, NM, BV, LT, CG, AR, TMF, LCP, CV, JD. Formal Analysis: OM, NM, BV, LT, CG, AR, JB, JAHM, LCP, JD. Funding Acquisition: CF, LMA, JMV, SK, SEL, MM, LF, KB, JB, DM, TMF, JAHM, LCP, CV, JD. Interpretation: OM, NM, BV, CF, LMA, RMW, XL, JMV, SK, FD, SEL, MM, LF, KB, JB, DM, TMF, JAHM, LCP, CV, JD. Investigation: OM, NM, BV, JB, DM, TMF, JAHM, LCP, CV, JD. Methodology: OM, NM, BV, RMW, XL, JMV, FD, MM, LF, KB, JB, DM, TMF, JAHM, LCP, CV, JD. Patient And Public Involvement: MM, LF, KB. Resources: OM, NM, BV, JAHM, LCP, JD. Software: OM, NM, BV, JB, DM, JAHM, LCP, JD. Supervision: LAC, FD, JB, DM, TMF, JAHM, LCP, CV, JD. Validation: OM, NM, BV, TMF, JAHM, LCP, CV, JD. Visualization: OM, NM, BV, TMF, JAHM, LCP, CV, JD. Writing – Original Draft: OM, NM, BV, TMF, JAHM, LCP, CV, JD. Writing – Review & Editing: OM, BV, LAC, CF, LMA, RMW, XL, JMV, SK, FD, SEL, KB, JB, DM, TMF, JAHM, LCP, CV, JD.

## Data sharing statement

We cannot make individual-level data available. Researchers can apply to UK Biobank (https://www.ukbiobank.ac.uk/enable-your-research/), CPRD (https://www.cprd.com/research-applications), and SIDIAP (https://www.sidiap.org/index.php/en/solicituds-en). We have made our

diagnostic code lists, code and results available on our GitHub (https://github.com/GEMINI-multimorbidity/) site and Shiny website (https://gemini-multimorbidity.shinyapps.io/atlas/). GWAS summary statistics will be available following acceptance at the GWAS Catalog (https://www.ebi.ac.uk/gwas/home).

## References

1 Chowdhury SR, Chandra Das D, Sunna TC, Beyene J, Hossain A. Global and regional prevalence of multimorbidity in the adult population in community settings: a systematic review and meta-analysis. *EClinicalMedicine* 2023; **57**: 101860.

2 Barnett K, Mercer SW, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *The Lancet* 2012; **380**: 37–43.

3 Carrasco-Ribelles LA, Roso-Llorach A, Cabrera-Bean M, *et al.* Dynamics of multimorbidity and frailty, and their contribution to mortality, nursing home and home care need: A primary care cohort of 1 456 052 ageing people. *EClinicalMedicine* 2022; **52**: 101610.

4 Violán C, Foguet-Boreu Q, Roso-Llorach A, *et al.* Burden of multimorbidity, socioeconomic status and use of health services across stages of life in urban areas: a cross-sectional study. *BMC Public Health* 2014; **14**: 530.

5 Cassell A, Edwards D, Harshfield A, *et al.* The epidemiology of multimorbidity in primary care: a retrospective cohort study. *British Journal of General Practice* 2018; **68**: e245–51.

6 Tran PB, Kazibwe J, Nikolaidis GF, Linnosmaa I, Rijken M, van Olmen J. Costs of multimorbidity: a systematic review and meta-analyses. *BMC Med* 2022; **20**: 234.

7 Ho IS-S, Azcoaga-Lorenzo A, Akbari A, *et al.* Examining variation in the measurement of multimorbidity in research: a systematic review of 566 studies. *Lancet Public Health* 2021; **6**: e587–97.

8 Nichols L, Taverner T, Crowe F, *et al.* In simulated data and health records, latent class analysis was the optimum multimorbidity clustering algorithm. *J Clin Epidemiol* 2022; **152**: 164–75.

9 Zhu Y, Edwards D, Mant J, Payne RA, Kiddle S. Characteristics, service use and mortality of clusters of multimorbid patients in England: a population-based study. *BMC Med* 2020; **18**: 78.

10 Koller D, Pathak GA, Wendt FR, *et al.* Epidemiologic and Genetic Associations of Endometriosis With Depression, Anxiety, and Eating Disorders. *JAMA Netw Open* 2023; **6**: e2251214.

11 Bulik-Sullivan B, Finucane HK, Anttila V, *et al.* An atlas of genetic correlations across human diseases and traits. *Nat Genet* 2015; **47**: 1236–41.

12 Bulik-Sullivan BK, Loh P-R, Finucane HK, *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 2015; **47**: 291–5.

13 Wolf A, Dedman D, Campbell J, *et al.* Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *Int J Epidemiol* 2019; **48**: 1740–1740g.

14 Recalde M, Rodríguez C, Burn E, *et al.* Data Resource Profile: The Information System for Research in Primary Care (SIDIAP). *Int J Epidemiol* 2022; **51**: e324–36.

15 Sudlow C, Gallacher J, Allen N, *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med* 2015; **12**: e1001779.

16 Kurki MI, Karjalainen J, Palta P, *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* 2023; **613**: 508–18.

17 Sollis E, Mosaku A, Abid A, *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res* 2023; **51**: D977–85.

18 Calderón-Larrañaga A, Vetrano DL, Onder G, *et al.* Assessing and Measuring Chronic Multimorbidity in the Older Population: A Proposal for Its Operationalization. *J Gerontol A Biol Sci Med Sci* 2017; **72**: 1417–23.

19 Kuan V, Denaxas S, Gonzalez-Izquierdo A, *et al.* A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. *Lancet Digit Health* 2019; **1**: e63–77.

20 Mbatchou J, Barnard L, Backman J, *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet* 2021; **53**: 1097–103.

21 Viechtbauer W. Conducting Meta-Analyses in R with the metafor Package. *J Stat Softw* 2010; **36**. DOI:10.18637/jss.v036.i03.

22 Mägi R, Morris AP. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics* 2010; **11**: 288.

23 GEMINI. GEMINI - Get Involved. 2021. https://sites.exeter.ac.uk/gemini/get-involved/ (accessed Jan 20, 2024).

24 Sodini SM, Kemper KE, Wray NR, Trzaskowski M. Comparison of Genotypic and Phenotypic Correlations: Cheverud's Conjecture in Humans. *Genetics* 2018; **209**: 941–8.

25 Anter E, Jessup M, Callans DJ. Atrial Fibrillation and Heart Failure. *Circulation* 2009; **119**: 2516–25.

26 Defeudis G, Mazzilli R, Tenuta M, *et al.* Erectile dysfunction and diabetes: A melting pot of circumstances and treatments. *Diabetes Metab Res Rev* 2022; **38**. DOI:10.1002/dmrr.3494.

27 Nirwan JS, Hasan SS, Babar Z-U-D, Conway BR, Ghori MU. Global Prevalence and Risk Factors of Gastro-oesophageal Reflux Disease (GORD): Systematic Review with Meta-analysis. *Sci Rep* 2020; **10**: 5814.

28 Smith JD, Irwin RW, Wolff ET. Two Unique Cases of Ciprofloxacin-Associated Avulsion of Ligament and Tendon. *Am J Phys Med Rehabil* 2018; **97**: e33–6.

29 Jeste D V., Gladsjo JA, Lindamer LA, Lacro JP. Medical Comorbidity in Schizophrenia. *Schizophr Bull* 1996; **22**: 413–30.

30 Momen NC, Plana-Ripoll O, Agerbo E, *et al.* Association between Mental Disorders and Subsequent Medical Conditions. *New England Journal of Medicine* 2020; **382**: 1721–31.

31 Crump C, Winkleby MA, Sundquist K, Sundquist J. Comorbidities and Mortality in Persons With Schizophrenia: A Swedish National Cohort Study. *American Journal of Psychiatry* 2013; **170**: 324–33.

32 Kugathasan P, Stubbs B, Aagaard J, Jensen SE, Munk Laursen T, Nielsen RE. Increased mortality from somatic multimorbidity in patients with schizophrenia: a Danish nationwide cohort study. *Acta Psychiatr Scand* 2019; **140**: 340–8.

33 Tan GSQ, Sloan EK, Lambert P, Kirkpatrick CMJ, Ilomäki J. Drug repurposing using real-world data. *Drug Discov Today* 2023; **28**: 103422.

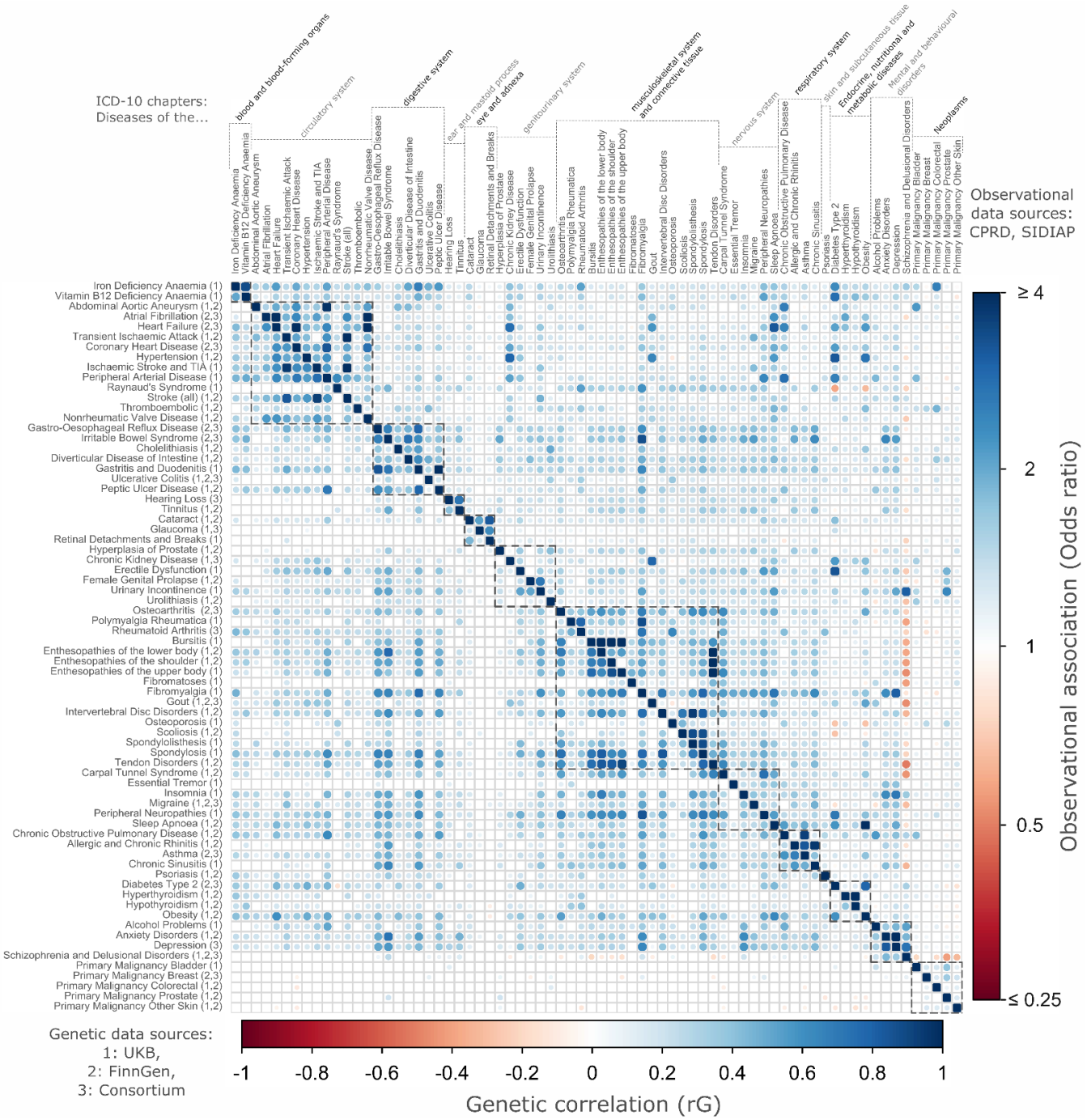
34 Benson RA, Meecham L, Fisher O, Loftus IM. Ultrasound screening for abdominal aortic aneurysm: current practice, challenges and controversies. *Br J Radiol* 2018; **91**: 20170306.

35 Ritz E, Haxsen V. Diabetic nephropathy and anaemia. *Eur J Clin Invest* 2005; **35**: 66–74.

36 Aroda VR, Edelstein SL, Goldberg RB, *et al.* Long-term Metformin Use and Vitamin B12 Deficiency in the Diabetes Prevention Program Outcomes Study. *J Clin Endocrinol Metab* 2016; **101**: 1754–61.

37 Masoli JAH, Pilling LC, Frayling TM. Genomics and multimorbidity. *Age Ageing* 2022; **51**. DOI:10.1093/ageing/afac285.

**Figure 1. Pairwise associations between 72 long-term chronic conditions: co-occurrence in observational data (upper right panel) and genetic correlation (lower left panel)**



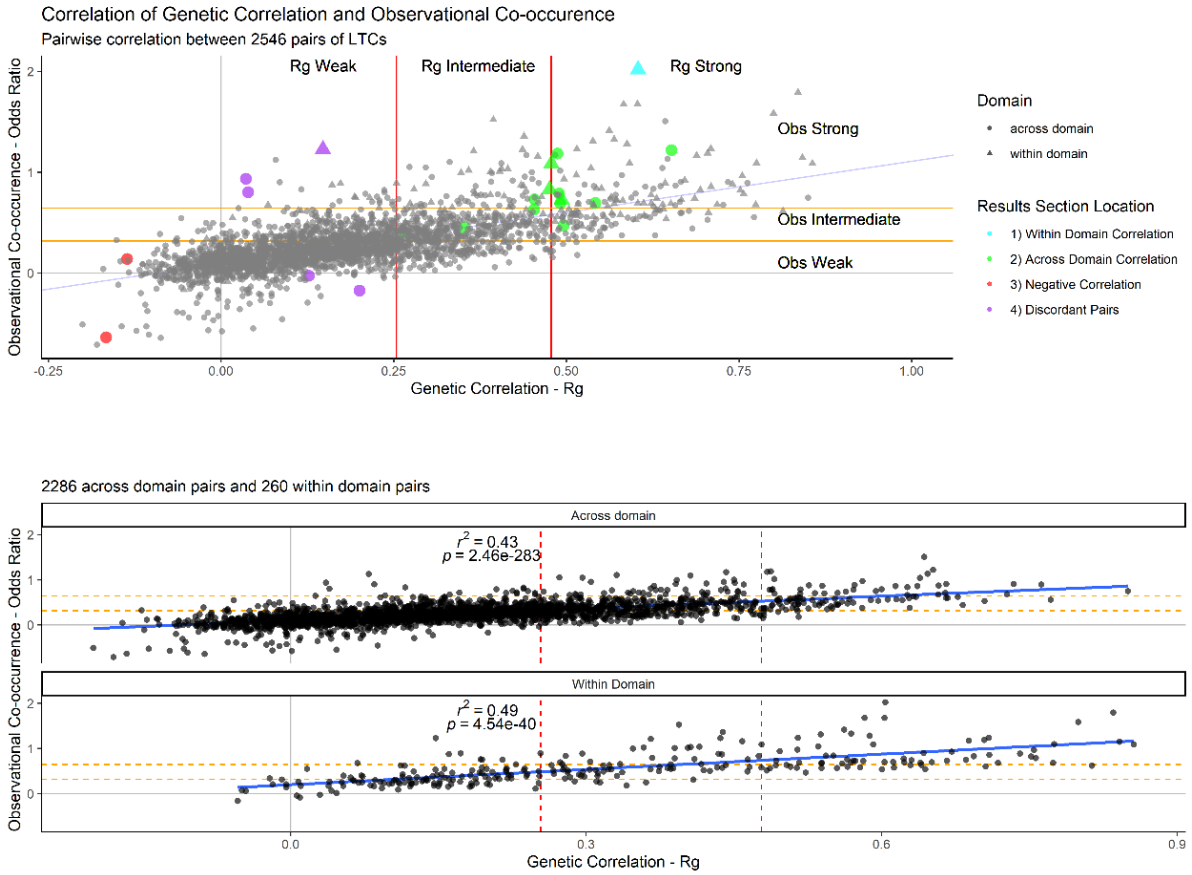

LTCs tested are common (≥0.5% prevalence in individuals aged 65 and older) and heritable (SNP-based heritability z-score >4) – see Methods. Observational associations (upper right triangle of heatmap) are Odds Ratios from meta-analysed logistic regression models from two population-representative primary care cohorts (CPRD and SIDIAP) comprised of individuals aged 65 or older and alive on January 1st, 2020 (to aid in readability, ORs >4 are set to 4, and ORs <0·25 are set to 0·25). Genetic correlations (lower right triangle of heatmap) are from meta-analysed GWAS summary statistics from up to three sources (UKB, FinnGen, and published consortium studies). LTCs within the same ICD-10 chapter (i.e., "within-domain") are highlighted with dotted lines. See Supplementary Tables 3 and 4 for full results, and web app for interactive version: https://gemini-multimorbidity.shinyapps.io/atlas.

**Table 1. Observational co-occurrence and genetic correlation between LTC pairs**

| All | Observational (primary care) associations | | | | |
|---|---|---|---|---|---|
| Genetic Correlation | Not Significant | Weak | Intermediate | Strong | Total |
| Not Significant | 100 (10·8) | 765 (82·7) | 53 (5·7) | 7 (0·8) | 925 (36·3) |
| Weak | 18 (2·1) | 692 (79·8) | 138 (15·9) | 19 (2·2) | 867 (34·1) |
| Intermediate | 0 (0) | 265 (45·8) | 245 (42·3) | 69 (11·9) | 579 (22·7) |
| Strong | 0 (0) | 18 (10·3) | 69 (39·4) | 88 (50·3) | 175 (6·9) |
| Obs Total | 118 (4·6) | 1740 (68·4) | 505 (19·8) | 183 (7·2) | 2546 (100) |
| | | | | | |
| Across-domain Pairs | Not Significant | Weak | Intermediate | Strong | Total |
| Not Significant | 98(4·3) | 727(31·8) | 45(2) | 6(0·3) | 876 (38·3) |
| Weak | 16(0·7) | 653(28·6) | 115(5) | 12(0·5) | 796 (34·8) |
| Intermediate | 0 (0) | 256(11·2) | 209(9·1) | 44(1·9) | 509 (22·3) |
| Strong | 0 (0) | 18(0·8) | 51(2·2) | 36(1·6) | 105 (4·6) |
| Obs Total | 114 (4·9) | 1654 (72·4) | 420 (18·4) | 98 (4·3) | 2286 (100) |
| | | | | | |
| Within-domain pairs | Not Significant | Weak | Intermediate | Strong | Total |
| Not Significant | 2(0·8) | 38(14·6) | 8(3·1) | 1(0·4) | 49 (18·9) |
| Weak | 2(0·8) | 39(15·0) | 23(8·8) | 7(2·7) | 71 (27·3) |
| Intermediate | 0 (0) | 9(3·5) | 36(13·8) | 25(9·6) | 70 (26·9) |
| Strong | 0 (0) | 0 | 18(6·9) | 52(20) | 70 (26·9) |
| Obs Total | 4 (1·5) | 86 (33·1) | 85(32·7) | 85(32·7) | 260 (100) |

For each pair the log-odds ratio of the least prevalent disease explained by the more prevalent disease adjusting for age and sex was meta-analysed across CPRD and SIDIAP. Observed LTC pairs were divided into terciles based on within-domain pairs; weak (Odds-Ratio [OR] ≤1·46) intermediate (1·46<OR≤1·90) and strong (OR>1·90). The genetic correlations of LTC pairs were divided into terciles based on within-domain pairs: weak ($R_g$≤0·25), intermediate (0·25<$R_g$≤0·48) and strong ($R_g$<0·48) correlation.

**Figure 2. The relationship between LTC observational co-occurrence and genetic correlation**




Scatter plots of the relationship between observed co-occurrence and genetic correlation of LTC pairs. (blue) linear regression line, (yellow) terciles of genetic correlation, (red) terciles of likelihood of observed co-occurrence. Terciles estimated based on within-domain LTC pairs. The upper panel shows all pairs with pairs discussed in result section 1:4 highlighted. The lower two panels show the across-domain (left, n=2,286) and within-domain (right, n=260) pairs. See Supplementary Tables 3 and 4 for full results, Supplementary Table 5 for the highlighted pairs and the web app for interactive versions: https://gemini-multimorbidity.shinyapps.io/atlas. See Supplementary Figure 3 for plot stratified by domain.