




# BMJ Open Protocol for the development of the Wales Multimorbidity e-Cohort (WMC): data sources and methods to construct a population-based research platform to investigate multimorbidity

Jane Lyons <sup>1</sup>, Ashley Akbari <sup>1</sup>, Utkarsh Agrawal,<sup>2</sup> Gill Harper,<sup>3</sup> Amaya Azcoaga-Lorenzo,<sup>2</sup> Rowena Bailey,<sup>1</sup> James Rafferty,<sup>1</sup> Alan Watkins <sup>1</sup>, Richard Fry <sup>1</sup>, Colin McCowan,<sup>2</sup> Carol Dezateux,<sup>3</sup> John P Robson,<sup>3</sup> Niels Peek,<sup>4</sup> Chris Holmes,<sup>5</sup> Spiros Denaxas,<sup>6</sup> Rhiannon Owen,<sup>7</sup> Keith R Abrams,<sup>7</sup> Ann John <sup>1</sup>, Dermot O'Reilly,<sup>8</sup> Sylvia Richardson,<sup>9</sup> Marlous Hall,<sup>10</sup> Chris P Gale,<sup>10</sup> Jan Davies,<sup>11</sup> Chris Davies,<sup>11</sup> Lynsey Cross,<sup>1</sup> John Gallacher,<sup>12</sup> James Chess <sup>13</sup>, Anthony J Brookes,<sup>14</sup> Ronan A Lyons <sup>1</sup>

**To cite:** Lyons J, Akbari A, Agrawal U, *et al.* Protocol for the development of the Wales Multimorbidity e-Cohort (WMC): data sources and methods to construct a population-based research platform to investigate multimorbidity. *BMJ Open* 2021;**11**:e047101. doi:10.1136/bmjopen-2020-047101

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2020-047101>).

Received 18 November 2020  
Revised 22 December 2020  
Accepted 06 January 2021



© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY. Published by BMJ.

For numbered affiliations see end of article.

## Correspondence to

Jane Lyons;  
j.lyons@swansea.ac.uk

## ABSTRACT

**Introduction** Multimorbidity is widely recognised as the presence of two or more concurrent long-term conditions, yet remains a poorly understood global issue despite increasing in prevalence.

We have created the Wales Multimorbidity e-Cohort (WMC) to provide an accessible research ready data asset to further the understanding of multimorbidity. Our objectives are to create a platform to support research which would help to understand prevalence, trajectories and determinants in multimorbidity, characterise clusters that lead to highest burden on individuals and healthcare services, and evaluate and provide new multimorbidity phenotypes and algorithms to the National Health Service and research communities to support prevention, healthcare planning and the management of individuals with multimorbidity.

**Methods and analysis** The WMC has been created and derived from multisourced demographic, administrative and electronic health record data relating to the Welsh population in the Secure Anonymised Information Linkage (SAIL) Databank. The WMC consists of 2.9 million people alive and living in Wales on the 1 January 2000 with follow-up until 31 December 2019, Welsh residency break or death. Published comorbidity indices and phenotype code lists will be used to measure and conceptualise multimorbidity.

Study outcomes will include: (1) a description of multimorbidity using published data phenotype algorithms/ontologies, (2) investigation of the associations between baseline demographic factors and multimorbidity, (3) identification of temporal trajectories of clusters of conditions and multimorbidity and (4) investigation of multimorbidity clusters with poor outcomes such as mortality and high healthcare service utilisation.

**Ethics and dissemination** The SAIL Databank independent Information Governance Review Panel has

## Strengths and limitations of this study

- Creation and access to a multisourced population based, deeply phenotyped e-cohort.
- Future use of this resource removes the need for data management and cleaning of source data, accelerating research and which could also support efforts for reproducibility of results.
- Variety of individual and household level data available on demography, health status, healthcare utilisation, both primary and secondary healthcare, and mortality to support a wide range of analytical approaches to addressing scientific questions.
- Input from multiple disciplines and institutions from across all four nations of the UK to help understand, measure and address multimorbidity.
- Routine data do not capture data on some important aspects, such as quality of life.

approved this study (SAIL Project: 0911). Study findings will be presented to policy groups, public meetings, national and international conferences, and published in peer-reviewed journals.

## INTRODUCTION

Multimorbidity is defined by the UK's Academy of Medical Sciences (AMS) and World Health Organization (WHO) as the presence of two or more concurrent long-term conditions, which is a global and growing phenomenon.<sup>1 2</sup> Multimorbidity is more prevalent in older individuals and associated with high healthcare utilisation and mortality, but with large numbers of patients of all age suffering from multimorbidity.<sup>3-6</sup> With an ageing population, it is estimated that two in

**Box 1 The Academy of Medical Sciences identified research gaps**

- ▶ The scale and nature of multimorbidity and how it is changing over time.
- ▶ Which clusters of conditions cause the biggest problems for patients.
- ▶ The causes of the most common clusters including links with sex, ethnicity, income and lifestyle.
- ▶ The best ways to prevent the patients developing multimorbidity, and whether this requires different approaches to just preventing individual conditions.
- ▶ How doctors can increase the benefits and reduce the risks of treatment for patients with multimorbidity.
- ▶ How to organise healthcare systems to deal with multimorbidity more effectively and how best to use digital technology in caring for patients.

three people in England aged 65 years or over will experience multimorbidity by 2035 and nearly one fifth will have complex multimorbidity (four or more conditions).<sup>7</sup>

Much of what is known about multimorbidity is based on a limited and fragmented knowledge base, largely derived from studies of older people in high-income countries or hospital populations.<sup>18</sup> The 2018 AMS report concluded that multimorbidity is an unhelpful term implying random assortment of disease when it often refers to clusters of specific diseases. Once identified, these disease clusters can be addressed specifically through research, healthcare policy development and service delivery.<sup>19</sup> The identification of previously unrecognised disease clusters may also provide biological and clinical insights into their aetiology, prevention and treatment. The AMS report identified specific research gaps and proposed a list of priorities (box 1). Several can be addressed through a combination of health data science, epidemiology and statistics and by exploiting the potential from creating deeply phenotyped cohorts from population and clinical data sources.

Responding to this agenda, we created a privacy protecting total population electronic cohort—the Wales Multimorbidity e-Cohort (WMC)—as a platform to study these issues in depth, collaborating with scientists from many different institutions and disciplines, clinicians, and members of the public from across the UK to create a broader team science approach.

The objectives of this work are to understand prevalence, trajectories and determinants of multimorbidity, and identify clusters causing the greatest healthcare burden. The WMC will also contribute data on incidence, prevalence and burden to the Global Burden of Diseases (GBD) Study,<sup>10 11</sup> and provide new multimorbidity phenotypes to e-cohorts with local participants, and phenotyping algorithms to many e-cohorts that use routine data.<sup>12</sup>

We expect that findings from these analyses will provide evidence to health policy leads in order to support prevention and the complex healthcare planning and management of multimorbid individuals. Members of the public are embedded in the research team to ensure the resource focuses on issues of concern to the public.

This paper describes the creation of the WMC and the statistical approaches that will be developed to support the diverse research objectives.

**METHODS**

The WMC was developed by linking multiple routinely collected population and clinical data sources on the population of Wales from 2000 to 2019. We used the privacy-protecting Secure Anonymised Information Linkage (SAIL) Databank, to contribute to the Health Data Research UK National Implementation Multimorbidity Resource (HNIMR) project and extended to 2020 for the Medical Research Council (MRC) funded Welsh Multimorbidity Machine Learning project.<sup>13 14</sup> SAIL is one of the most comprehensive, privacy protecting, linked data Trusted Research Environments in the UK. SAIL uses data from many different sources and provides linkage at individual and household level.<sup>15</sup> It has supported many different study designs, including large-scale community-based or clinical condition-based observational studies, disease surveillance, evaluation of natural experiments of environmental interventions, embedded trials and the Dementias Platform UK.<sup>16–23</sup>

**Cohort design and characteristics**

The WMC is a clearly defined complete population cohort. Cohort entry includes all residents in Wales, alive and living on 1 January 2000. Cohort censorship was defined by the first date of migration out of Wales/residency break, death, or the study endpoint on 31 December 2019 (figure 1). Within these constraints, the cohort is designed to be without selection bias and to achieve complete follow-up. WMC also provides a fully generalisable population sample against which findings from more selected samples may be compared.

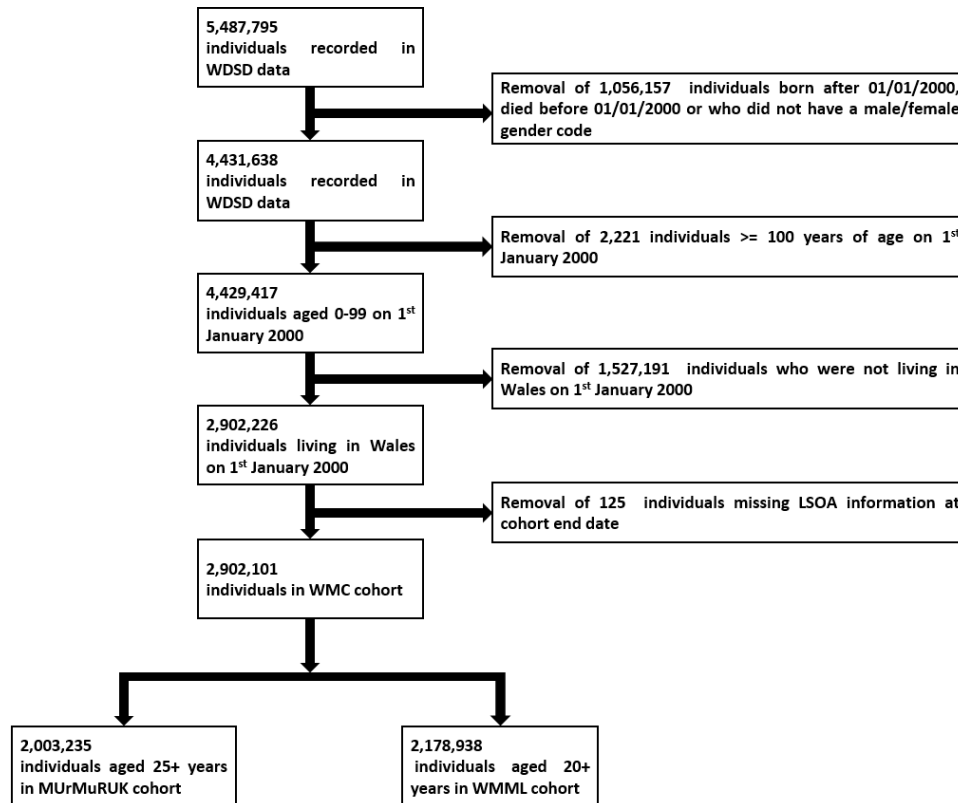
The WMC contains 2 902 101 individuals aged 0–99 at cohort start date with 46 million person years of follow-up available (table 1, figures 2 and 3, online supplemental appendix table A1 and A2). Individuals have a minimum of 1-day follow-up (cohort end date = 2 January 2000) and maximum of 20 years of follow-up (cohort end date = 31 December 2019).

The Heatmap in figure 3 visualises the person years of follow-up by age, sex and area level deprivation. The more years of follow-up available the darker the colour. Age is calculated at the cohort start, therefore, younger individuals will have more years of available follow-up compared with older individuals. On average, there are less person years of follow-up available for the least deprived 15–24 years old compared with their respective age group in other areas of Wales.

**Data sources**

The WMC has used and combined anonymised health, social and environmental data held within the SAIL Databank ([www.saildatabank.com](http://www.saildatabank.com)).

The baseline characteristics for the WMC have been created using the Welsh Demographic Service Dataset



**Figure 1** WMC flow diagram, based on inclusion criteria. LSOA; lower layer super output area, WSD; Welsh Demographic Service Dataset, WMC; Wales Multimorbidity e-Cohort, WMML; Welsh Multimorbidity Machine Learning.

(WSD) and the Annual District Death Extract (ADDE) mortality registry data from the Office for National Statistics. The WSD contains administrative information concerning the resident population of Wales that

are registered to a Welsh General Practice, a free to use National Health Service (NHS) system at the point of primary care registration in the UK. The ADDE data contains information about the dates and causes of all deaths relating to residents in Wales, including those that died outside of Wales. SAIL holds general practitioner (GP) data for approximately 80% of the population with coverage extending to all local authorities in Wales. The Welsh Longitudinal General Practice data will be used to identify the subpopulation of individuals who are registered to a practice providing data to SAIL to identify which individuals have GP data present and avoid under-estimation of conditions or severity of conditions not managed through hospital admission.

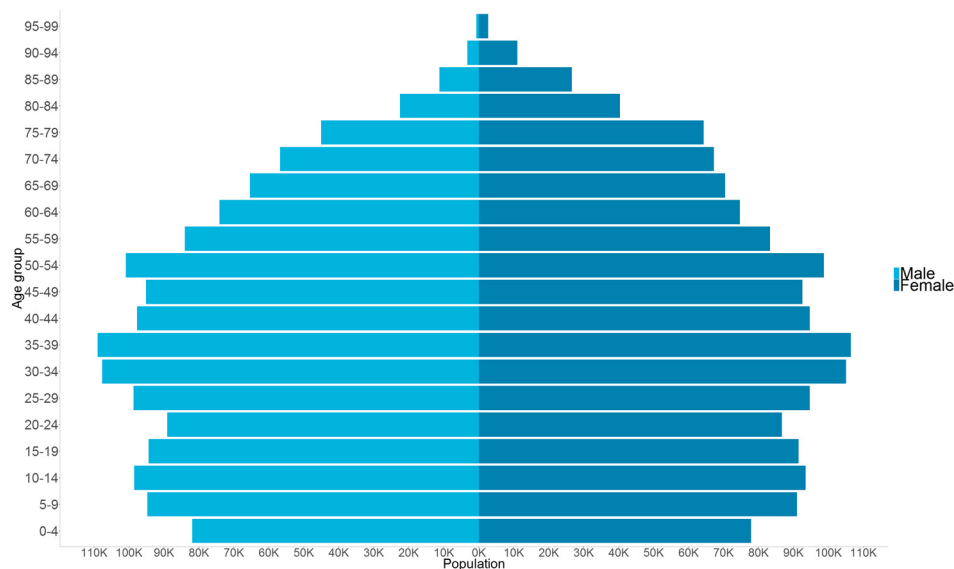
The Welsh Health Survey Dataset and the National Survey for Wales Dataset with data on well-being measures, social class, education, housing and wealth are available for 9905 and 33 295 cohort participants respectively.<sup>24</sup>

### Anonymised linkage fields

Linkage fields are used to anonymously link between data sources in the SAIL Databank and have been previously described elsewhere.<sup>13 14 25</sup> SAIL uses a multiple encryption system in which a trusted third party, the NHS Wales Informatics Service, uniquely matches identities (NHS number, name, date of birth and residential address/Unique Property Reference Number (UPRN)) and replaces these with unique identifiers. For individuals this is called an Anonymised Linkage Field (ALF) and

Table 1 WMC baseline demographics	
WMC characteristics	n (%)
Cohort size	2 902 101 (100)
Full coverage (1 January 2000–31 December 2019)	1 714 484 (59.08)
Residency break/emigration	643 472 (22.17)
Mortality	544 145 (18.75)
Primary care data available	2 470 874 (85.14)
Care home residency at cohort end	97 006 (3.34)
Mean age in years (range) at cohort start	39 (0–99)
Sex	
Female	1 472 113 (50.70)
Male	1 429 988 (49.30)
WIMD 2011 Quintile at cohort start	
1. Most deprived	605 203 (20.85)
2	589 479 (20.31)
3	584 039 (20.12)
4	557 319 (19.20)
5. Least deprived	566 061 (19.51)

WIMD, Welsh Index of Multiple Deprivation; WMC, Wales Multimorbidity e-Cohort.



**Figure 2** WMC pyramid for age (years) at cohort inception. WMC, Wales Multimorbidity e-Cohort.

Residential Anonymised Linkage Field (RALF) for pseudonymised residences before uploading data to SAIL.

**Demographic data**

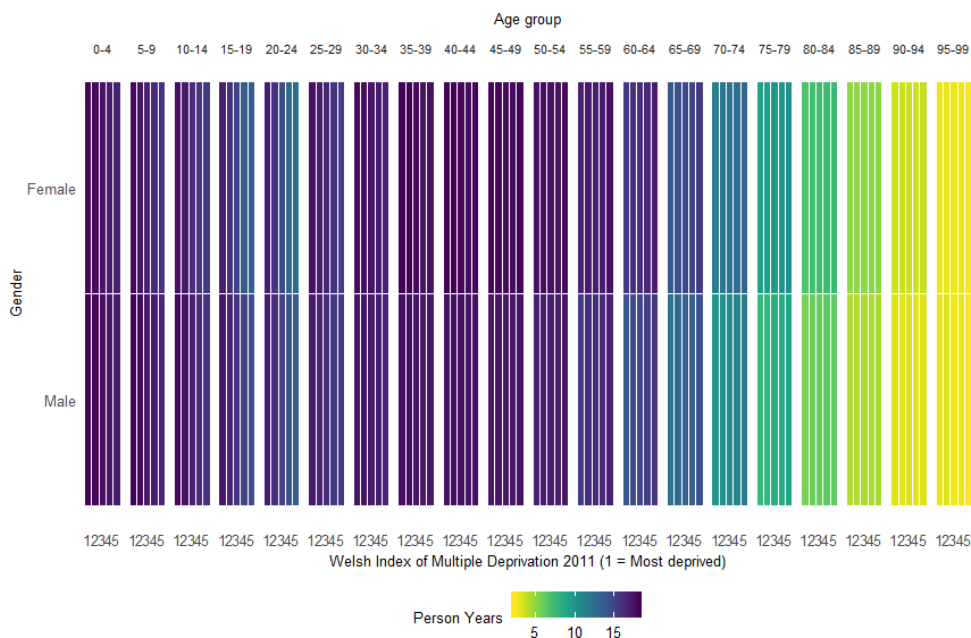
The cohort includes the following variables: ALF, age in years, sex, date of death, date of movement out of Wales, RALF at both cohort inception and cohort end and Care Home Anonymised Linkage Fields (CHALFs) at cohort end date. The CHALF was derived from a data extract from Care Inspectorate Wales in 2020 for all adult care home settings.<sup>18</sup> Geographical variables associated with the RALF and CHALF include Lower layer Super Output Area (LSOA) 2001 at cohort inception and LSOA 2011 at cohort end. These have been mapped to the Welsh Index

of Multiple Deprivation version 2011 and 2019, respectively, to derive socioeconomic deprivation quintiles and urban/rurality categories.<sup>26 27</sup>

**Health data**

All admissions to hospital (inclusive of critical care admissions), outpatient, emergency department attendances treated in NHS hospitals as well as disease registries and laboratory test results data are available for cohort participants, GP data for diagnoses and treatments from SAIL providing practices are data for approximately 80% of the population.<sup>28</sup>

All relevant health events recorded in clinical data sources will be joined onto the WMC to identify diagnosis



**Figure 3** Heatmap of person years of WMC follow-up, by age group, sex and area-level deprivation at cohort inception. WMC, Wales Multimorbidity e-Cohort.



**Table 2** Clinical data sources available for the WMC

Data source	Period covered	No and percentage of WMC individuals with data
Critical Care Data Set	01-Jan-2007–31-Dec-2019	79 521 (2.7)
Welsh Cancer Incidence Surveillance Unit	01-Jan-2000–31-Dec-2016	328 792 (11.3)
Welsh Results Reporting Services	01-Jan-2015–10-Dec-2018	1 540 754 (53.1)
Emergency Department Data Set	01-Apr-2009–31-Dec-2019	1 579 665 (54.4)
Patient Episode Database for Wales)	01-Jan-2000–31-Dec-2019	2 129 384 (73.4)
Out Patient Dataset for Wales	01-Apr-2004–31-Dec-2019	2 177 081 (75.0)
Welsh Longitudinal General Practice	01-Jan-2000–31-Dec-2019	2 400 313 (82.7)

Please note clinical data sources will be updated on a monthly/quarterly basis. WMC, Wales Multimorbidity e-Cohort.

of conditions, treatments and various significant health events that occur across multisourced linked health data per person (table 2 and figure 4).

The Upset plot in figure 4 demonstrates the number of WMC participants that have interacted with the various healthcare settings from 1 January 2000 to their cohort censorship end date.<sup>29</sup> For example, 780 830 (26.9%) individuals have used GP, inpatient, outpatient and emergency department services as well as had at least one laboratory test within their WMC coverage.

### Phenotyping the e-cohort

Published comorbidity indices and phenotype code lists (International Classification of Diseases 10th revision (ICD-10), OPCS Classification of Interventions and Procedures version 4 (OPCS4) and primary care Read Codes version 2) will be used to measure and conceptualise multimorbidity. These include those created by: CALIBER initiative; Charlson Comorbidity Index; Common Mental Disorders (CMD); Elixhauser Comorbidity Index; GBD Study and the NHS Quality and Outcomes Framework (QOF).<sup>30–41</sup> Diagnostic codes relating to HIV will not be included in any outputs to conform with SAIL policies. They are part of the list of redacted codes not allowed to be used for research using the data.<sup>42</sup> All ICD-10 and

OPCS4 codes provided at the three character level were expanded to include all children terms.

### CALIBER

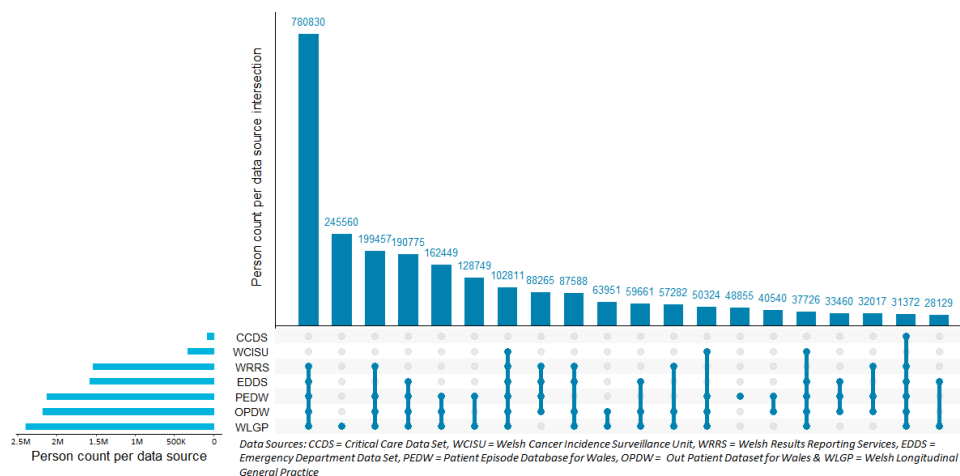
Phenotyping algorithms created from the CALIBER resource using ICD-10, OPCS4 and Read Codes will be used to identify 300 physical and mental health conditions recorded in both primary and secondary healthcare.<sup>31 39</sup>

There are 1645 distinct ICD-10 codes (at three and four-character level) for 300 conditions, however, when capturing all ICD-10 codes to include variation in coding entry (eg, C796– instead of C796) and expanding the code list to the four-character level (F200 instead of F20), there are 3702 distinct ICD-10 codes (at the four-character level) recorded in the inpatient data. This is important to note as to link solely on standardised codes would result in loss of information and potential reporting of false negatives.

There are 587 distinct OPCS4 codes (at three and four-character level) for 28 conditions and 8588 distinct Read Codes (at the five-character level) for 275 conditions.

### Charlson Comorbidity Index

The Aylin and Bottle Charlson amended ICD-10 code list will be used for inpatient diagnosis and the Metcalfe *et*



**Figure 4** Number of WMC individuals utilising healthcare services recorded in multisource data sources, 20 most common combinations presented. WMC, Wales Multimorbidity e-Cohort.

*et al*<sup>33</sup> Charlson Read Code list will be utilised for primary care recorded diagnosis.<sup>32 33</sup>

The ICD-10 codes have been taken from the pool of diagnosis codes recorded within hospital admissions data, containing 1024 distinct codes (at the four-character level) for 16 conditions. The GP data contains 4545 distinct Read Codes at the five-character level.

#### Common mental disorders

The John *et al* validated algorithm will be used to identify CMD in GP data.<sup>30 40 41</sup> The algorithm has used a combination of diagnosis, treatment and symptoms Read Codes in identifying CMD. Individuals with CMD are identified as either having a historical diagnosis code, currently treated or, having a current diagnosis/current symptom code. There are 89 distinct diagnosis codes, 15 symptom codes and 601 treatment codes.

#### Elixhauser Comorbidity Index

The Quan *et al* (2005) Elixhauser ICD-10 code list will be utilised for inpatient diagnosis and the Metcalfe *et al*<sup>33</sup> Elixhauser Read Code list will be utilised for primary care recorded diagnosis.<sup>33 34</sup>

The ICD-10 codes have been taken from the pool of diagnosis codes recorded within hospital admissions data and contains 1423 distinct codes (at the four-character level) for 30 conditions. The general practice data contains 6074 distinct Read codes at the five-character level.

#### GBD Study

The GBD 2019 ICD-10 codes will be used to identify 130 health conditions in secondary healthcare data. There are 3497 distinct ICD-10 codes at the three and four-character level.<sup>38</sup>

#### Quality Outcome Framework

The QOF conditions business rule V.38 will be used to identify 18 health conditions in primary care data.<sup>35</sup> The 18 conditions are asthma, atrial fibrillation, obesity, coronary heart disease, chronic obstructive pulmonary disease, cancer, chronic kidney disease, dementia, depression, diabetes, epilepsy, heart failure, hypertension, learning difficulties, peripheral arterial disease, rheumatoid arthritis, serious mental illness and stroke. There are 2275 distinct Read Codes available at the five-character level for the 18 QOF conditions.

#### Statistical analysis

The WMC provides an accessible research ready data asset to further understanding of multimorbidity through the use of biostatistical and machine learning approaches. Our collaborative team will work across a number of projects to develop and evaluate statistical and machine learning algorithms to address the following broad analytical challenges:

- ▶ What is the prevalence of multimorbidity in the WMC, and how does prevalence of multimorbidity change over time?

- ▶ What are common clusters of multimorbidity in the WMC, and how do they correspond to or differ from, common clusters of multimorbidity identified in other datasets?
- ▶ Which clusters of multimorbidity occur less frequently than one would expect based on the prevalence of their constituent conditions?
- ▶ How does multimorbidity develop across the life course (ie, trajectories)?
- ▶ What are the biological, psychological and social determinants of different clusters and trajectories of multimorbidity?
- ▶ Which clusters and trajectories of multimorbidity are associated with poor health outcomes?
- ▶ Which clusters and trajectories of multimorbidity are associated with high service utilisation?
- ▶ Does multimorbidity in specific groups (eg, patients with musculoskeletal conditions) differ from multimorbidity in general?

The overarching aim is to evaluate and provide new multimorbidity phenotypes and algorithms to the NHS and research communities to support prevention, health-care planning and the management of individuals with multimorbidity.

We will draw on both methods from statistics (eg, regression analysis, longitudinal mixed models, multiple correspondence analysis, factor analysis,<sup>43</sup> multistate models and latent class analysis) and machine learning (eg, k-means clustering, semantic similarity clustering, market basket analysis, network models<sup>44</sup> and deep learning). We will use resampling methods to assess the stability of identified multimorbidity clusters and develop visualisation techniques to summarise multimorbidity clusters and their associations with risk factors and outcomes.

Analyses will be coded in R, WinBUGS, and Python and made available to WMC users via a Git library to maximise transparency and reproducibility.<sup>45</sup>

#### Patient and public involvement

The proposal to develop WMC was submitted to the independent Information Governance Review Panel (IGRP) that includes members of the public (IGRP Project: 0911). We worked with this group to refine the study protocol. The scientific steering group includes two members of the public who have contributed to this paper. The HNIMR has a work package on patient and public involvement with a panel drawn from across the UK which meets to discuss the research work and feed into the research and dissemination plans.

#### ETHICS AND DISSEMINATION

The use of deidentified data in SAIL complies with National Research Ethics Service (NRES) guidance.<sup>46</sup> Applications to use data held within the SAIL Databank, an ISO: 27001 and UK Statistics Authority (UKSA) Digital Economy Act (DEA) accredited Trusted Research Environment, must first be approved by the independent

IGRP. This panel contains individuals with expertise in data governance and protection, including the Chair of the Wales NRES Committee, Caldicott Guardians and members of the public. WMC was approved by IGRP on 26 June 2019.

Findings from this study will be disseminated widely through a variety of routes, including to health policy and NHS leads across UK, the AMS and the Royal Colleges, as well as traditional scientific outlets. The team includes NHS clinicians and informaticians to allow for early NHS adoption of useful findings. Members of the public embedded in the team will create plain English summaries and lead at public facing meetings.

#### Author affiliations

<sup>1</sup>Population Data Science, Swansea University Medical School, Swansea, UK

<sup>2</sup>School of Medicine, University of St Andrews, St Andrews, UK

<sup>3</sup>Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK

<sup>4</sup>Health e-Research Centre, Institute of Population Health, University of Manchester, Manchester, UK

<sup>5</sup>Department of Statistics, Oxford University, Oxford, UK

<sup>6</sup>Institute of Health Informatics, University College London, London, UK

<sup>7</sup>Department of Health Sciences, University of Leicester, Leicester, UK

<sup>8</sup>Epidemiology and Public Health, Queens University Belfast, Belfast, UK

<sup>9</sup>Department of Epidemiology and Public Health, MRC Biostatistics Unit, Cambridge, UK

<sup>10</sup>School of Medicine, University of Leeds, Leeds, UK

<sup>11</sup>Members of the public, Swansea, UK

<sup>12</sup>Department of Psychiatry, Oxford University, Oxford, UK

<sup>13</sup>Renal Unit, Swansea Bay University Health Board, Swansea, UK

<sup>14</sup>Department of Genetics, University of Leicester, Leicester, UK

**Twitter** Ashley Akbari @AshleyAkbari, Richard Fry @richfry and Ann John @ProfAnnJohn

**Acknowledgements** This study makes use of anonymised data held in the Secure Anonymised Information Linkage (SAIL) Databank. We would like to acknowledge all the data providers and people who make anonymised data available for research. The authors would like to extend their gratitude and acknowledgement to the NHS, the SAIL Consumer panel as well as the IGRP who approved this project.

**Contributors** Conceptualisation of study JL, AA, UA, GH, CM, DOR and RAL; data curation and analysis JL; original draft writing JL, review and editing of manuscript JL, AA, UA, GH, AA-L, RB, JR, AW, RF, CM, CD, JPR, NP, CH, SD, RO, KRA, AJ, DOR, SR, MH, CPG, JD, CD, LC, JG, JC, AJB and RAL.

**Funding** Two UK-wide collaborative efforts have been formed to address several of the AMS report priorities: Measuring and Understanding Multimorbidity using Routine Data in the UK – (MURMURUK) and Application of machine learning to discover new multimorbidity phenotypes associated with poorer outcomes (WMLL). This work was supported by Health Data Research UK (HDR-9006; CFC0110) and the Medical Research Council (MR/S027750/1). Health Data Research UK is funded by: UK Medical Research Council; Engineering and Physical Sciences Research Council; Economic and Social Research Council; National Institute for Health Research (England); Chief Scientist Office of the Scottish Government Health and Social Care Directorates; Health and Social Care Research and Development Division (Welsh Government); Public Health Agency (Northern Ireland); British Heart Foundation and Wellcome Trust.

**Disclaimer** The views and opinions expressed therein are those of the authors and do not necessarily reflect those of the funding agencies, NHS organisations or Welsh Government.

**Competing interests** None declared.

**Patient consent for publication** Not required.

**Provenance and peer review** Not commissioned; peer reviewed for ethical and funding approval prior to submission.

**Supplemental material** This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

**Open access** This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: <https://creativecommons.org/licenses/by/4.0/>.

#### ORCID iDs

Jane Lyons <http://orcid.org/0000-0002-4407-770X>

Ashley Akbari <http://orcid.org/0000-0003-0814-0801>

Alan Watkins <http://orcid.org/0000-0003-3804-1943>

Richard Fry <http://orcid.org/0000-0002-7968-6679>

Ann John <http://orcid.org/0000-0002-5657-6995>

James Chess <http://orcid.org/0000-0001-8805-6962>

Ronan A Lyons <http://orcid.org/0000-0001-5225-000X>

#### REFERENCES

- 1 The Academy of Medical Sciences. Multimorbidity: a priority for global health research, 2018. Available: <https://acmedsci.ac.uk/file-download/82222577>
- 2 WHO. The challenges of a changing world. the world health report 2008—primary health care (now more than ever), 2008. Available: <https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0237186&type=printable>
- 3 Barnett K, Mercer SW, Norbury M, *et al*. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *Lancet* 2012;380:37–43.
- 4 Cassell A, Edwards D, Harshfield A, *et al*. The epidemiology of multimorbidity in primary care: a retrospective cohort study. *Br J Gen Pract* 2018;68:e245–51.
- 5 Nunes BP, Flores TR, Mielke GI, *et al*. Multimorbidity and mortality in older adults: a systematic review and meta-analysis. *Arch Gerontol Geriatr* 2016;67:130–8.
- 6 Hall M, Donno TB, Yan AT, *et al*. Multimorbidity and survival for patients with acute myocardial infarction in England and Wales: latent class analysis of a nationwide population-based cohort. *PLoS Med* 2018;15:e1002501.
- 7 Kingston A, Robinson L, Booth H, *et al*. Projections of multimorbidity in the older population in England to 2035: estimates from the population ageing and care simulation (PACSim) model. *Age Ageing* 2018;47:374–80.
- 8 Diederichs C, Berger K, Bartels DB. The measurement of multiple chronic diseases—a systematic review on existing multimorbidity indices. *J Gerontol A Biol Sci Med Sci* 2011;66:301–11.
- 9 Ford JC, Ford JA. Multimorbidity: will it stand the test of time? *Age Ageing* 2018;47:6–8.
- 10 Vos T, Flaxman AD, Naghavi M, *et al*. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the global burden of disease study 2010. *Lancet* 2012;380:2163–96.
- 11 Steel N, Ford J, Newton J. Mortality, causes of death, years of life lost, years lived with a disability, and disability-adjusted life years in the countries of the UK and 150 English local authority areas 1990–2016: a systematic analysis for the global burden of disease study 2016. *Lancet* 2018.
- 12 Bauermeister S, Orton C, Thompson S, *et al*. The dementias platform UK (DPUK) data portal. *Eur J Epidemiol* 2020;35:601–11.
- 13 Lyons RA, Jones KH, John G, *et al*. The Sail databank: linking multiple health and social care datasets. *BMC Med Inform Decis Mak* 2009;9:3.
- 14 Ford DV, Jones KH, Verplancke J-P, *et al*. The Sail Databank: building a national architecture for e-health research and evaluation. *BMC Health Serv Res* 2009;9:157.
- 15 Lyons RA, Ford DV, Moore L, *et al*. Use of data linkage to measure the population health effect of non-health-care interventions. *Lancet* 2014;383:1517–9.





- 16 Lyons RA, Turner S, Lyons J, *et al.* All Wales injury surveillance system revised: development of a population-based system to evaluate single-level and multilevel interventions. *Inj Prev* 2016;22 Suppl 1:i50–5.
- 17 Snooks HA, Anthony R, Chatters R, *et al.* Support and assessment for fall emergency referrals (safer) 2: a cluster randomised trial and systematic review of clinical effectiveness and cost-effectiveness of new protocols for emergency ambulance paramedics to assess older people following a fall with referral to community-based care when appropriate. *Health Technol Assess* 2017;21:1–218.
- 18 Hollinghurst J, Akbari A, Fry R, Hollingsworth JP, Rodgers SE, *et al.* Study protocol for investigating the impact of community home modification services on hospital utilisation for fall injuries: a controlled longitudinal study using data linkage. *BMJ Open* 2018;8:e026290.
- 19 Mizen A, Song J, Fry R, *et al.* Longitudinal access and exposure to green-blue spaces and individual-level mental health and well-being: protocol for a longitudinal, population-wide record-linked natural experiment. *BMJ Open* 2019;9:e027289.
- 20 Szakmany T, Walters AM, Pugh R, *et al.* Risk factors for 1-year mortality and hospital utilization patterns in critical care survivors: a retrospective, observational, population-based data linkage study. *Crit Care Med* 2019;47:15–22.
- 21 Rodgers SE, Bailey R, Johnson R, *et al.* Emergency hospital admissions associated with a non-randomised housing intervention meeting national housing quality Standards: a longitudinal data linkage study. *J Epidemiol Community Health* 2018;72:896–903.
- 22 Paranjothy S, Evans A, Bandyopadhyay A, *et al.* Risk of emergency hospital admission in children associated with mental disorders and alcohol misuse in the household: an electronic birth cohort study. *Lancet Public Health* 2018;3:e279–88.
- 23 Schnier C, Wilkinson T, Akbari A, *et al.* The secure Anonymised information linkage databank dementia e-cohort (SAIL-DeC). *Int J Popul Data Sci* 2020;5:1121.
- 24 Welsh Government. Discontinuities in results for health-related lifestyle and general health between the Welsh health survey and national survey for Wales. Available: <https://gov.wales/sites/default/files/statistics-and-research/2019-02/discontinuities-results-health-related-lifestyle-general-health-between-welsh-health-survey-national-survey-wales-2018.pdf> [Accessed 24 Aug 2020].
- 25 Rodgers SE, Lyons RA, Dsilva R, *et al.* Residential anonymous linking fields (RALFs): a novel information infrastructure to study the interaction between the environment and individuals' health. *J Public Health* 2009;31:582–8.
- 26 Welsh Government. Welsh index multiple deprivation index. Available: <https://gov.wales/welsh-index-multiple-deprivation-index-guidance> [Accessed 9 Apr 2020].
- 27 Office for National Statistics. 2011 rural/urban classifications. Available: <https://www.ons.gov.uk/methodology/geography/geographicalproducts/ruralurbanclassifications/2011ruralurbanclassification> [Accessed 9 Apr 2020].
- 28 Thayer D, Rees A, Kennedy J, *et al.* Measuring follow-up time in routinely-collected health datasets: challenges and solutions. *PLoS One* 2020;15:e0228545.
- 29 CRAN.R. Nils Gehlenborg (2019). UpSetR: a more scalable alternative to Venn and Euler diagrams for visualizing intersecting sets. R package version 1.4.0. Available: <https://CRAN.R-project.org/package=UpSetR>
- 30 John A, McGregor J, Fone D, *et al.* Case-Finding for common mental disorders of anxiety and depression in primary care: an external validation of routinely collected data. *BMC Med Inform Decis Mak* 2016;16:35.
- 31 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.
- 32 Bottle A, Aylin P. Comorbidity scores for administrative data benefited from adaptation to local coding and diagnostic practices. *J Clin Epidemiol* 2011;64:1426–33.
- 33 Metcalfe D, Masters J, Delmestri A, *et al.* Coding algorithms for defining Charlson and Elixhauser co-morbidities in Read-coded databases. *BMC Med Res Methodol* 2019;19:115.
- 34 Quan H, Sundararajan V, Halfon P, *et al.* Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care* 2005;43:1130–9.
- 35 NHS. Quality and outcomes framework (QOF) business rules V 38 2017-2018 October code release. Available: <https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/quality-and-outcomes-framework-qof/quality-and-outcome-framework-qof-business-rules/quality-and-outcomes-framework-qof-business-rules-v-38-2017-2018-october-code-release> [Accessed 21 Aug 2019].
- 36 Charlson ME, Pompei P, Ales KL, *et al.* A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis* 1987;40:373–83.
- 37 Elixhauser A, Steiner C, Harris DR, *et al.* Comorbidity measures for use with administrative data. *Med Care* 1998;36:8–27.
- 38 Global Burden of Disease Collaborative Network. Global burden of disease study 2017 (GBD 2017) causes of death and nonfatal causes mapped to ICD codes. Seattle, United States of America: Institute for health metrics and evaluation (IHME), 2018. Available: <http://ghdx.healthdata.org/record/ihme-data/gbd-2017-cause-icd-code-mappings> [Accessed 01 Jun 2020].
- 39 Denaxas S, Gonzalez-Izquierdo A, Direk K, *et al.* UK phenomics platform for developing and validating electronic health record phenotypes: caliber. *J Am Med Inform Assoc* 2019;26:1545–59.
- 40 John A, DelPozo-Banos M, Gunnell D, *et al.* Contacts with primary and secondary healthcare prior to suicide: case-control whole-population-based study using person-level linked routine data in Wales, UK, 2000-2017. *Br J Psychiatry* 2020;217:717–24.
- 41 Ware JE, Gandek B. Overview of the SF-36 health survey and the International quality of life assessment (IQOLA) project. *J Clin Epidemiol* 1998;51:903–12.
- 42 Citizenspace.com. Legally unsharable clinical codes - NHS Digital - Citizen Space [Internet]. Available: <https://nhs-digital.citizenspace.com/standards-assurance/legally-unsharable-clinical-codes> [Accessed 9 Nov 2020].
- 43 Pages J. *Multiple factor analysis by example using R* [Internet]. Philadelphia, PA: Chapman & Hall/CRC, 2014.
- 44 Marx P, Antal P, Bolgar B, *et al.* Comorbidities in the diseasome are more apparent than real: what Bayesian filtering reveals about the comorbidities of depression. *PLoS Comput Biol* 2017;13:e1005487.
- 45 Lunn DJ, Thomas A, Best N, *et al.* WinBUGS — a Bayesian modelling framework: concepts, structure, and extensibility. *Stat Comput* 2000;10:325–37.
- 46 Jones KH, Ford DV, Jones C, *et al.* A case study of the secure anonymous information linkage (Sail) gateway: a privacy-protecting remote access system for health-related research and evaluation. *J Biomed Inform* 2014;50:196–204.



## Appendix

Table A1: WMC participants categorised by age group and sex at cohort start

Age group	Sex	Count	Percentage
00-04	Male	81,915	2.82
00-04	Female	77,873	2.68
05-09	Male	94,737	3.26
05-09	Female	90,940	3.13
10-14	Male	98,466	3.39
10-14	Female	93,447	3.22
15-19	Male	94,345	3.25
15-19	Female	91,440	3.15
20-24	Male	89,037	3.07
20-24	Female	86,666	2.99
25-29	Male	98,622	3.40
25-29	Female	94,592	3.26
30-34	Male	107,671	3.71
30-34	Female	104,986	3.62
35-39	Male	108,964	3.75
35-39	Female	106,312	3.66
40-44	Male	97,637	3.36
40-44	Female	94,599	3.26
45-49	Male	95,071	3.28
45-49	Female	92,478	3.19
50-54	Male	100,866	3.48
50-54	Female	98,606	3.40
55-59	Male	83,949	2.89
55-59	Female	83,210	2.87
60-64	Male	74,115	2.55
60-64	Female	74,591	2.57
65-69	Male	65,354	2.25
65-69	Female	70,389	2.43
70-74	Male	56,746	1.96
70-74	Female	67,227	2.32
75-79	Male	45,027	1.55
75-79	Female	64,274	2.21
80-84	Male	22,441	0.77
80-84	Female	40,344	1.39
85-89	Male	11,184	0.39
85-89	Female	26,540	0.91
90-94	Male	3,208	0.11
90-94	Female	10,951	0.38
95-99	Male	633	0.02
95-99	Female	2,648	0.09

Table A2: WMC average person years of follow up, categorised by age group, sex and WIMD 2011 at cohort start

Age group	Sex	WIMD quintiles	2011	Average Pys
00-04	Male		1	18.16
00-04	Male		2	17.93
00-04	Male		3	17.63
00-04	Male		4	17.28
00-04	Male		5	17.11
05-09	Male		1	18.06
05-09	Male		2	17.81
05-09	Male		3	17.26
05-09	Male		4	16.91
05-09	Male		5	16.50
10-14	Male		1	17.93
10-14	Male		2	17.46
10-14	Male		3	16.76
10-14	Male		4	16.24
10-14	Male		5	15.98
15-19	Male		1	17.30
15-19	Male		2	16.73
15-19	Male		3	15.75
15-19	Male		4	14.89
15-19	Male		5	14.34
20-24	Male		1	16.96
20-24	Male		2	16.04
20-24	Male		3	15.29
20-24	Male		4	14.03
20-24	Male		5	13.59
25-29	Male		1	17.18
25-29	Male		2	16.71
25-29	Male		3	16.22
25-29	Male		4	15.57
25-29	Male		5	15.47
30-34	Male		1	17.44
30-34	Male		2	17.41
30-34	Male		3	17.11
30-34	Male		4	16.82
30-34	Male		5	16.60
35-39	Male		1	17.66
35-39	Male		2	17.63
35-39	Male		3	17.41
35-39	Male		4	17.27

35-39	Male	5	17.22
40-44	Male	1	17.50
40-44	Male	2	17.63
40-44	Male	3	17.55
40-44	Male	4	17.40
40-44	Male	5	17.57
45-49	Male	1	17.23
45-49	Male	2	17.39
45-49	Male	3	17.28
45-49	Male	4	17.24
45-49	Male	5	17.48
50-54	Male	1	16.68
50-54	Male	2	16.96
50-54	Male	3	16.89
50-54	Male	4	16.91
50-54	Male	5	17.30
55-59	Male	1	15.46
55-59	Male	2	16.03
55-59	Male	3	16.20
55-59	Male	4	16.31
55-59	Male	5	16.78
60-64	Male	1	14.07
60-64	Male	2	14.64
60-64	Male	3	14.96
60-64	Male	4	15.19
60-64	Male	5	15.80
65-69	Male	1	12.14
65-69	Male	2	12.70
65-69	Male	3	13.24
65-69	Male	4	13.46
65-69	Male	5	14.21
70-74	Male	1	9.73
70-74	Male	2	10.23
70-74	Male	3	10.59
70-74	Male	4	11.02
70-74	Male	5	11.58
75-79	Male	1	7.46
75-79	Male	2	7.73
75-79	Male	3	8.14
75-79	Male	4	8.29
75-79	Male	5	8.79
80-84	Male	1	5.55
80-84	Male	2	5.82
80-84	Male	3	6.02

80-84	Male	4	6.24
80-84	Male	5	6.30
85-89	Male	1	4.20
85-89	Male	2	4.18
85-89	Male	3	4.28
85-89	Male	4	4.27
85-89	Male	5	4.41
90-94	Male	1	3.09
90-94	Male	2	3.07
90-94	Male	3	3.22
90-94	Male	4	2.95
90-94	Male	5	3.13
95-99	Male	1	2.87
95-99	Male	2	3.19
95-99	Male	3	2.64
95-99	Male	4	2.77
95-99	Male	5	2.32
00-04	Female	1	18.03
00-04	Female	2	17.82
00-04	Female	3	17.37
00-04	Female	4	16.99
00-04	Female	5	16.73
05-09	Female	1	17.84
05-09	Female	2	17.44
05-09	Female	3	16.89
05-09	Female	4	16.27
05-09	Female	5	15.92
10-14	Female	1	17.57
10-14	Female	2	17.09
10-14	Female	3	16.20
10-14	Female	4	15.60
10-14	Female	5	15.51
15-19	Female	1	16.93
15-19	Female	2	16.08
15-19	Female	3	14.88
15-19	Female	4	13.68
15-19	Female	5	13.21
20-24	Female	1	16.94
20-24	Female	2	15.81
20-24	Female	3	14.61
20-24	Female	4	12.89
20-24	Female	5	12.48
25-29	Female	1	17.53
25-29	Female	2	16.99



25-29	Female	3	16.47
25-29	Female	4	15.77
25-29	Female	5	15.34
30-34	Female	1	17.94
30-34	Female	2	17.77
30-34	Female	3	17.39
30-34	Female	4	16.94
30-34	Female	5	16.77
35-39	Female	1	18.18
35-39	Female	2	18.12
35-39	Female	3	17.78
35-39	Female	4	17.47
35-39	Female	5	17.49
40-44	Female	1	18.11
40-44	Female	2	18.16
40-44	Female	3	17.91
40-44	Female	4	17.71
40-44	Female	5	17.92
45-49	Female	1	17.92
45-49	Female	2	17.93
45-49	Female	3	17.82
45-49	Female	4	17.66
45-49	Female	5	17.97
50-54	Female	1	17.49
50-54	Female	2	17.69
50-54	Female	3	17.49
50-54	Female	4	17.44
50-54	Female	5	17.87
55-59	Female	1	16.79
55-59	Female	2	17.09
55-59	Female	3	17.00
55-59	Female	4	17.06
55-59	Female	5	17.54
60-64	Female	1	15.53
60-64	Female	2	16.04
60-64	Female	3	16.23
60-64	Female	4	16.28
60-64	Female	5	16.96
65-69	Female	1	13.76
65-69	Female	2	14.28
65-69	Female	3	14.70
65-69	Female	4	14.92
65-69	Female	5	15.52
70-74	Female	1	11.43

70-74	Female	2	11.93
70-74	Female	3	12.27
70-74	Female	4	12.57
70-74	Female	5	13.12
75-79	Female	1	9.06
75-79	Female	2	9.35
75-79	Female	3	9.70
75-79	Female	4	9.82
75-79	Female	5	10.25
80-84	Female	1	6.78
80-84	Female	2	7.07
80-84	Female	3	7.19
80-84	Female	4	7.26
80-84	Female	5	7.50
85-89	Female	1	4.98
85-89	Female	2	5.02
85-89	Female	3	4.97
85-89	Female	4	5.07
85-89	Female	5	5.07
90-94	Female	1	3.45
90-94	Female	2	3.53
90-94	Female	3	3.61
90-94	Female	4	3.55
90-94	Female	5	3.66
95-99	Female	1	2.66
95-99	Female	2	2.87
95-99	Female	3	2.70
95-99	Female	4	2.75
95-99	Female	5	2.48