UNIVERSITY OF CALIFORNIA SAN DIEGO

SAN DIEGO STATE UNIVERSITY

Advanced Epidemiology Methods for the Study of Infectious Diseases: Examples using COVID-19 and HIV

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

in

Public Health (Epidemiology)

by

Jonathan Davitte

Committee in charge:

San Diego State University

Professor Richard Shaffer, Chair
Professor Stephanie Brodine

University of California San Diego

Professor Natasha Martin
Professor Heather Pines
Professor Rany Salem

2021

The Dissertation of Jonathan Davitte is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____
Chair

University of California San Diego

San Diego State University

2021

DEDICATION

For all of the patience, support, and love throughout this complicated 5-year journey, this Dissertation is dedicated to my husband Beau. From our beginning in San Diego, to the suburbs of Philadelphia, to lockdown in London, and finally to new roots in the Pennsylvania forests; you have always been the foundation on which this work could be built.

I'd also like to acknowledge the incredible, on-going work of the U.S. Department of Defense HIV/AIDS Prevention Program (DHAPP); and, especially the U.S. Embassy Staff in Mozambique and the *Forças Armadas de Defesa de Moçambique*. I am lucky to have worked with such an outstanding organization. I am the researcher I am today because of DHAPP and our partner militaries. While natural disasters, funding, and a pandemic ultimately prevented HIV in Mozambique being the sole focus of my dissertation; the skills that enabled the research presented in my Dissertation are a product of my work at DHAPP and its partner militaries.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Richard Shaffer for his support as the chair of my committee. He saw the potential for me to continue my education and encouraged me to embark on this journey. I now proudly can call myself an 'Epidemiologist' because of his support and encouragement. While we no longer have the pleasure of working together professionally, I count myself lucky that I have been able to continue to collaborate with the incredible DHAPP team.

I would like to acknowledge both Dr. Bonnie Tran and Antonio Langa. Dr. Tran and Mr. Langa have been essential in continuing to push forward the 'Treat All' study, despite two hurricanes, funding gaps, and a global pandemic. Chapter 5 of this dissertation would not be possible without their tireless work and dedication.

I would also like to acknowledge the Human Genetics team at GlaxoSmithKline for their support in allowing me to use UK Biobank data when my original dissertation plans fell through in the middle of the COVID-19 pandemic. Without their incredible support, I would not have been able to complete this dissertation.

Finally, I would like to acknowledge the UK Biobank participants for their dedication to participating in ongoing research and electronic health record linkage. This work and the incredible work of other UK Biobank researchers would not have been possible without their dedication to science. All UK Biobank data was accessed in accordance with GlaxoSmithKline's UK Biobank Application #20361.

Chapter 2, in full, is currently being prepared for submission for publication of the material. Davitte, J.; Pines, H.; Martin, N.; Salem, R.; Brodine, S.; Shaffer, R. The dissertation author was the primary investigator and author of this material.

Chapter 3, in full, is currently being prepared for submission for publication of the material. Davitte, J.; Salem, R.; Pines, H.; Martin, N.; Brodine, S.; Shaffer, R. The dissertation author was the primary investigator and author of this material.

Chapter 4, in full, is currently being prepared for submission and publication of the material. Davitte, J.; Pines, H.; Martin, N.; Brodine, S.; Salem, R.; Shaffer, R. The dissertation author was the primary investigator and author of this material.

VITA

2007	Bachelor of Arts, University of Alabama at Birmingham
2010	Master of Public Health, University of Alabama at Birmingham
2021	Doctor of Philosophy, University of California San Diego and San Diego State University

ABSTRACT OF THE DISSERTATION

Advanced Epidemiology Methods for the Study of Infectious Diseases:
Examples using COVID-19 and HIV

by

Jonathan Davitte

Doctor of Philosophy in Public Health (Epidemiology)

University of California San Diego, 2021
San Diego State University, 2021

Professor Richard Shaffer, Chair

The COVID-19 epidemic has highlighted a number of important challenges for infectious disease Epidemiologic research: 1) scaling causal-inference efforts across the human disease phenome; 2) understanding the long-term consequences of a novel disease without robust longitudinal data; and, 3) leveraging non-traditional data types for infectious disease research. Our dissertation provides three examples of advanced Epidemiologic methods that illustrate how researchers may address one or more of these challenges.

Given the prevalence of multiple comorbidities, interrelated disease states may represent a more complete picture of COVID-19 infection severity risk compared to a disease-by-disease approach. We

used a bias-adjusted, three-step Latent Class Analysis (LCA) method to identify patterns of comorbidities from 31 disease indicators; and, measured their relationship to severe COVID-19 infection among 176,894 participants in the UK Biobank cohort. We identified 5 distinct comorbidity patterns from 31 disease indicators, assessed using clinical diagnosis records from UK Biobank's comprehensive EHR data linkage between 2015-2019. Our results identified significantly increased risk for severe COVID-19 infection, with substantial heterogeneity in effect sizes, for each of our 4 comorbidity latent classes compared to our '*Healthy*' latent class.

We investigated the associations between genetic liability to severe COVID-19 infection, measured with Polygenic Risk Scores (PRS), and 31 comorbidity phenotypes derived from linked electronic health record (EHR) data in the past 20 years. PRS for very severe COVID-19 infection were associated with increased risk for uncomplicated diabetes, uncomplicated hypertension, obesity, and renal failure. Our research indicates that the same genetic composition that increases an individual's risk for COVID-19 may also influence their risk for other important comorbid diseases.

Proximity to military bases may be an indicator of accessibility to military sexual partners; and, help identify important local HIV epidemics. We estimated the relationship between travel time to the nearest military base and HIV-status among 7,514 young women recruited at local venues. Our study found that adolescent girls and young women that meet or congregate near military bases were at a slightly elevated risk for HIV-infection in the combined sample, but only in 1 of our 4 military bases in stratified analysis.

Chapter 1: Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS-COV-2 / COVID-19) has emerged at a time where Epidemiologists have unprecedented access to massive quantities of observational data, numerous types of health data, advanced methods enabled in open-source software, and powerful computing resources. Despite the opportunities presented by our access to data, methods to leverage these data, and tools to implement these methods, the COVID-19 epidemic has highlighted a number of important challenges for infectious disease Epidemiologic research: 1) scaling causal-inference efforts across the human disease phenome; 2) understanding the long-term consequences of a novel disease without robust longitudinal data; and, 3) leveraging non-traditional data types for infectious disease research. Each advanced Epidemiology method example in this dissertation will address one or more of these challenges; and, provide a potential roadmap for how Epidemiologists may respond to novel or emerging infectious diseases in the future.

Prior to the 1950s, researchers commonly used vital statistics to conduct cross-sectional and time series studies of noninfectious disease.[1] The lack of longitudinal data from these efforts limited Epidemiologic research focused on causal inference. Consequently, new funding enabled researchers to develop cohorts of individuals with extensive, active follow-up over long periods of time. However, these prospective studies have faced new challenges in the twenty-first century, namely declining research support and participating rates.[2] More recently, electronic health records (EHR) databases now provide Epidemiologists a low-cost method of quickly accessing incredibly rich, longitudinal data on large populations.[1] For example, the United Kingdom (UK) National Health Service (NHS) offers international researchers the opportunity to explore 'cradle to grave' longitudinal EHR data. The UK is one of the few countries that combines a single-payer-and-provider comprehensive healthcare system, free at the point of care, with extensive national data resources across the entire population of 65 million.[3] Consequently, Epidemiologists levering EHR data from the UK have incredible quantities of clinical diagnoses, laboratory results, and other routinely collected observational data across massive numbers of individuals over a very long period of time.

Another source of data, relevant to Epidemiologic research, that has continued to grow in recent years is genetic data. Biobanks have been established to study DNA or other molecular markers derived from peripheral circulated blood, epithelial cells from the inner cheek or mouth, blood cells from umbilical cords, urine samples, or diseased tissue.[3] Population-based biobanks, repositories consisting of biological tissue donated by thousands of individuals from the general population who may or may not have a specific disease, are increasingly being linked to EHR data. These population-based biobanks provide a powerful tool that enable Epidemiologists to understand the genetic and environmental determinants of disease. In the UK, national EHR data sources are continually being liked with population-based biobanks such as the 100,000 Genomes Project (also known as Genomics England[4]) and the UK Biobank[5], enabling rapid investigation of simple or complex disease across participant populations with diverse genetic backgrounds.[3] In response to the COVID-19 epidemic, UK Biobank is providing regular releases of diagnostic COVID-19 testing data, GP (primary care) data provided directly by the system suppliers, hospital inpatient data, critical care data, and mortality data to facilitate research into the determinants and consequences of COVID-19.[6]

Geographic information systems (GIS) data is another growing data source with specific relevance to the study of infectious disease. Advances in GIS technology have made it incredibly easy to connect spatially referenced physical and social phenomena to population patterns of health, disease, and well-being.[7] The spread of infectious disease is inherently a spatial process: and, applications of GIS data and methods can enable 1) improving infectious disease surveillance; 2) incorporating mobility data into infectious disease forecasting; 3) enabling digital contact tracing; 4) integrating geographic data into epidemiologic models; and, 5) investigating geographic social vulnerabilities and health disparities.[8]

Despite the quantities and types of data available to for infectious disease research, there is no clear consensus on how best to capitalize on these resources for robust causal inference research across the entire human disease phenome, quickly and efficiently in response to a novel disease. Causal inference research typically requires pre-existing knowledge (often derived from prior research) of the potential confounders that may impact exposure-outcome effect estimates in order to accurately quantify

the exposure-outcome relationship. However, as COVID-19 has illustrated, there may not be decades of research into important relationships between exposures, outcomes, and confounders immediately available for Epidemiologists to refine their causal inference research. New COVID-19 research, emerging daily, has highlighted two common approaches being used in current practice to quantify the relationship between various diseases (across the disease phenome) and COVID-19 outcomes: a '*simultaneous estimate*' approach where researchers evaluate a large number of exposures simultaneously in a single regression model to identify significant relationships with COVID-19; and, a '*disease-by-disease*' approach where researchers investigate a single exposure's relationship to COVID-19 outcomes.[9–14]

While '*simultaneous estimate*' efforts are valuable for hypothesis-generating (e.g. risk factors as 'predictor, without attention to cause' or 'covariate with a statistically significant association with the outcome'); they are frequently mistaken by consumers or misreported by authors as 'possible cause under investigation'. It is common practice for researchers to present adjusted associations for many diseases (exposures) and covariates (for confounding control) with their outcome from a single model in a single table, suggesting that all estimates can be interpreted similarly, if not identically, typically as a total-effect estimate.[15] However, the interpretation of an effect estimate may differ based on which variables are considered the 'exposure' and which variables are considered 'confounders'. Consequently, a causal model for one exposure may be entirely different from that of a completely different, but related exposure. This practice is referred to in Epidemiology as the 'Table 2 Fallacy'. The '*disease-by-disease*' approach, where researchers investigate a single exposure as a potential cause of COVID-19 outcomes, is difficult to scale quickly across the vast number of health conditions that may be relevant to COVID-19. In addition, pre-existing research that can inform robust causal inference for a given disease exposure will not exist in the presence of a novel disease. Given interrelated disease states, it can be incredibly difficult to identify potential disease confounders; or, sequence research appropriately to generate the required knowledge of these disease-confounder relationships. Methods employed in both the '*disease-by-disease*'

and the '*simultaneous estimate*' approaches often treat multiple diseases as distinct disease states; or, they simply count the number of diseases in each patient, assuming each disease has equal importance.

In response to the difficulty in '*scaling causal-inference efforts across the human disease phenome*', Latent Class Analysis (LCA) represents a potential method to inform targeting of future disease-by-disease causal inference. LCA uses indicators to probabilistically assign a population into subgroups with distinct profiles based on observed indicators. Using 'diseases' as indicators will result in categories that represent groups of individuals with different disease profiles. Researchers may then use additional methods to assess the relationship between latent class membership (e.g. specific disease patterns) and an outcome of interest to identify significant differences in the outcome by latent class membership.[16] The availability of EHR data lends itself to this method as researchers can easily generate numerous disease 'indicators' for the LCA; and, use the results to understand the specific disease patterns that are relevant to a given outcome.

Analyses leveraging genetic data may provide solutions for '*understanding the long-term consequences of a novel disease without robust longitudinal data*' as well as the '*non-traditional data types for infectious disease research*' issues. In the case of COVID-19, although the respiratory system is the primary site affected by the COVID-19 virus, infection has proven to be a major threat to other organ systems, including cardiovascular, gastrointestinal, renal, central nervous, and reproductive systems.[17] While there is the potential for long-term impact of COVID-19 across the disease-phenome, the absence of long-term follow-up post-COVID-19 infection prevents traditional Epidemiologic methods for investigating these effects. While genetic Epidemiologists have been investigating the impact of genetic variants across many diseases in recent history, the COVID-19 epidemic represents the first novel disease to emerge in the presence of large biobanks linked with EHR data. Consequently, COVID-19 represents a key opportunity in applying genetic analysis methods used to describe other diseases to a novel disease.

Typically the magnitude of effect and variance explained by a single genetic variant is small and individually have limited utility when evaluating genetic liability for a given trait and/or health outcome(s).[18–20] The 'Common Disease, Common Variant' hypothesis posits that genetic variants with

appreciable frequency in the population at large, but relatively low probability that variant carriers will express the disease (penetrance), are the major contributors to genetic susceptibility to common diseases.[21] The cumulative risk derived from aggregating contributions of the many common and uncommon variants associated with a complex trait or disease, such as COVID-19, is referred to as a polygenic risk score (PRS).[22] PRSs are commonly defined as the sum of trait genome-wide-associated (single nucleotide polymorphisms (SNPs) weighted by their effect sizes to provide an overall measure of an individual's genetic liability to that trait or disease.[23] Consequently, PRSs can achieve substantially greater predictive power for a given trait by including a larger number of SNPs in the PRS compared to restricting to only SNPs that reach GWAS genome-wide significance (e.g. $p < 5x10^{-8}$).[24] While PRSs have many applications, they have been used extensively to identify shared genetic etiology between two traits.[25,25–29] In identifying diseases with shared genetic risk to COVID-19, PRSs can inform hypotheses for further genetic causal inference efforts; and, identify health outcomes that warrant additional scrutiny once sufficient longitudinal data has accumulated.

Given that the spread of infectious disease is inherently a spatial process, the use of GIS data is key to '*leveraging non-traditional data types for infectious disease research*'. Similar to research using genetic data, there are a growing number of applications of GIS data and methods in infectious disease research as the scale and depth of GIS data accumulates. In research of the human immunodeficiency virus (HIV/AIDS), individuals are often reluctant to disclose sexual contacts or do not know their contact details. COVID-19 poses similar challenges for contact tracing, with many individuals hesitant to respond to requests for information from researchers.[30] The use of GIS data provides researchers the ability to investigate spatial clustering and mobility of populations important for the spread of infectious diseases; potentially highlighting interactions between important populations for disease transmission or locations with significant disease transmission.

Chapter 2 of this dissertation will identify specific patterns of comorbid disease and the relationship between these patterns and severe COVID-19 infection in the UK Biobank cohort. We hypothesize that there may be comorbidity latent classes that are associated with COVID-19 severity;

and, that the strength and/or direction of these associations varies among the latent classes. The use of our LCA method, leveraging the extensive EHR data in the UK Biobank, provides a real-world example that may address the '*scaling causal-inference efforts across the human disease phenome*' issue.

Chapter 3 of this dissertation will investigate whether there is shared genetic etiology between individual comorbid diseases and severe COVID-19 infection using PRS in the UK Biobank cohort. We hypothesized that increased polygenic risk for severe COVID-19 infection is positively associated with risk for some of these 31 comorbidities. This example is intended to address both the '*understanding the long-term consequences of a novel disease without robust longitudinal data*' as well as the '*leveraging non-traditional data types for infectious disease research*' issues, again leveraging the extensive EHR data in the UK Biobank with additional genetic data.

Chapter 4 of this dissertation will examine the association between proximity to the local military base (measured via travel time to the nearest base from AGYW recruitment location) and HIV infection among AGYW (15-35 years of age) living in communities surrounding military bases in Mozambique. We hypothesized that congregating or meeting sexual partners at venues in closer proximity to military bases is positively associated with HIV infection. This example is intended to address the '*leveraging non-traditional data types for infectious disease research*' issue through a novel use of GIS data in the context of HIV research.

Chapter 2: Comorbidity Patterns and COVID-19 Infection Severity: Latent Class Analysis of UK Biobank Electronic Health Records

2.1 Abstract

While new COVID-19 research is emerging daily, our knowledge of the relationships between comorbidities (pre-existing clinical conditions) on COVID-19 infection has largely been limited to: 1) a disease-by-disease approach where researchers investigate a given comorbidity as a potential cause of COVID-19 outcomes; or, 2) evaluating a large number of comorbidities simultaneously to identify significant relationships between a given comorbidity and COVID-19. Given the prevalence of multiple comorbidities, especially in older age, analyzing interrelated disease states may represent a more complete picture of COVID-19 infection severity risk compared to a disease-by-disease approach. We used a bias-adjusted, three-step Latent Class Analysis (LCA) method to identify distinct patterns of comorbidities, based on 31 disease indicators assessed in linked electronic health record (EHR) data, and measured their relationship to severe COVID-19 infection (hospitalization and/or death) among 170,734 participants in the UK Biobank cohort. We identified 5 distinct comorbidity patterns: all with significantly increased risk for severe COVID-19 infection compared to our '*Healthy*' latent class. Our research highlights the importance of considering patterns of comorbidities and their combined effects with respect to COVID-19 infection.

2.2 Introduction

As of February 2021, the severe acute respiratory syndrome coronavirus 2 (SARS-COV-2 / COVID-19) has now infected more than 109 million individuals globally, resulting in 2.4 million deaths.[31] The United Kingdom has had more than 4 million confirmed cases and currently has the highest number of COVID-19 deaths per 100,000 population worldwide (176.90 deaths / 100,000 population).[32]

While new COVID-19 research is emerging daily, our knowledge of the relationships between comorbidities (pre-existing clinical conditions[33]) on COVID-19 infection has largely been limited to: 1) a disease-by-disease approach where researchers investigate a given comorbidity as a potential cause of COVID-19 outcomes[9–11]; or, 2) evaluating a large number of comorbidities simultaneously to identify

significant relationships between a given comorbidity and COVID-19.[12–14] Given the prevalence of multiple comorbidities, especially in older age, consideration of interrelated disease states may represent a more complete picture of COVID-19 infection severity risk compared to a disease-by-disease approach.[34]

The 'disease-by-disease' approach, where researchers investigate a given comorbidity as a potential cause of COVID-19 outcomes, is difficult to scale quickly across the vast number of health conditions that may be relevant to COVID-19. Furthermore, research that establishes a given disease as a causal risk factor for COVID-19 may be difficult to contextualize in the presence of interrelated disease states. For example, recent research by Cao et.al. report 'obesity' as an independent risk factor for severe outcomes of COVID-19.[9] An alternative study by Gao et.al. identified that patients with hypertension had a two-fold increase in the relative risk of COVID-19 mortality compared to patients without hypertension.[10] Neither study investigates the combined effect of 'obesity' and 'hypertension' on COVID-19 outcomes, despite the common nature of comorbid hypertension and obesity, where forty percent of obese patients had hypertension in the Cao et.al. study.[9] Gao et.al. did not measure either body mass index or obesity as part of their investigation.[10]

A common alternative to the 'disease-by-disease' approach is to evaluate a large number of diseases simultaneously to identify the strongest relationships between a given disease and COVID-19 outcomes. These efforts are valuable as hypothesis-generating efforts (e.g. risk factors as 'predictor, without attention to cause' or 'covariate with a statistically significant association with the outcome'); but, are frequently mistaken by consumers or misreported by authors as 'possible cause under investigation'. While it is common practice for these research efforts to present adjusted associations for many comorbidities and covariates (for confounding control) with their COVID-19 outcome from a single model in a single table, this suggests that all estimates can be interpreted similarly, if not identically, typically as total-effect estimates (e.g. the 'Table 2 Fallacy').[15] However, the interpretation of an effect estimate may differ based on which variables are considered the 'exposure' and which variables are considered 'confounders'. Consequently, a causal model for one comorbidity may be entirely different from that of a completely different, but related disease. This practice is referred to in Epidemiology as the

'*Table 2 Fallacy*'. For example, a recent study of COVID-19 outcomes in a US integrated health system (Kaiser Permanente Southern California) sought to "disentangle the effect of BMI, associated comorbidities and medications, time, neighborhood-level income and education, and other factors on the risk for COVID-19". The authors reported adjusted relative risks for body mass index (BMI) category, age, sex, race/ethnicity, smoking status, time and 20 comorbidity indicators using a single multivariable Poisson regression.[12] Based on their findings, the authors conclude that 'we demonstrate the leading role severe obesity has over other highly correlated risk factors, providing a clear target for early intervention'. However, it is extremely likely that the causal model for 'organ transplant' on COVID-19 is entirely different from the intended model for a causal effect of BMI on COVID-19, despite both 'risk factors' being considered as 'possible causes under investigation'. While we have selected this particular study as particularly vulnerable to the '*Table 2 Fallacy*', there are many more examples of this issue in other recent published research, especially with respect to the impact of common comorbidities on COVID-19 outcomes.[12–14]

Methods employed in both the '*disease-by-disease*' and the '*simultaneous estimate*' approaches often treat multiple comorbidities as distinct disease states; or, they simply count the number of diseases in each patient, assuming each disease has equal importance. In the context of COVID-19, there has not been decades of research identifying which comorbidities or combinations of comorbidities are important for COVID-19 infection outcomes. Furthermore, the prevalence of multiple comorbidities increases with age, which may be particularly important given the large impact of COVID-19 on elderly populations worldwide.[35,36] One approach that may be used to identify patterns of comorbidities is Latent Class Analysis (LCA), where the study population is probabilistically assigned into latent classes with distinct disease profiles based on observed disease indicators. Additional methods may then be used to assess the strength, direction, and significance for the relationship between latent class membership (e.g. specific comorbidity patterns) and an outcome of interest.[16]

In our study, we used a bias-adjusted, three-step LCA method to identify distinct patterns of comorbidities, based on 31 disease indicators, and measured their relationship to severe COVID-19

infection in the UK Biobank cohort. We hypothesized that there may be comorbidity latent classes are associated with COVID-19 severity; and, that the strength of these associations will vary among the latent classes. The intention of our research is not to measure causal relationships between comorbidity latent classes and our COVID-19 outcomes. Our research is intended to describe important comorbidity patterns that may inform more targeted research on the impact of specific diseases and their relevant, interrelated disease states on COVID-19.

2.3 Materials and Methods

2.3.1 Data Sources

We used data from the UK Biobank, a prospective cohort study providing detailed characterization of over half a million UK-based persons aged 40-69 years at recruitment from 2006 to 2010, with continuous follow-up to present day through additional, bespoke data collection efforts as well as regular linkage to National Health Service (NHS) electronic health record (EHR) data and other registries (e.g. Cancer and Mortality).[5] Participants were assessed in 22 centers throughout the UK, providing socioeconomic and ethnic heterogeneity as well as urban-rural mix. Data collection at the baseline assessment visit included: electronic signed consent; a self-completed touch-screen questionnaire; brief computer-assisted interview; physical and functional measures; and collection of blood, urine, and saliva. In response to the COVID-19 epidemic, UK Biobank is providing regular releases of diagnostic COVID-19 testing data, GP (primary care) data provided directly by the system suppliers, hospital inpatient data, critical care data, and mortality data to facilitate research into the determinants and consequences of COVID-19.[6]

Our study leveraged the following data sources: baseline assessment data for demographics; hospital episodes for inpatient clinical diagnoses and critical care episodes, released on 22 February 2021, coded using the International Classification of Diseases, Tenth Revision (ICD-10); mortality events from the death registry, released on 16 February 2021; cancer diagnoses, coded in ICD-10, released in March 2019; and, primary care data supplied by 'The Phoenix Partnership' (TPP) system provider for clinical diagnoses coded using the Clinical Terms Version 3 (CTV3), released on 14 April 2021;. We did not

leverage primary care data supplied by the 'Egton Medical Information Systems' (EMIS) system provider given that clinical diagnoses in from 2015-2019 were primarily coded using SNOMED-CT terms for which a validated, published definition of the Elixhauser comorbidities using SNOMED-CT terms was not available at the time of this research.

Elixhauser Comorbidities

Comorbidity summary measures have been developed to help classify patients according to their overall disease burden. Elixhauser et.al defined a set of 31 comorbidity indicators, which have been translated by Quan et.al. for use in administrative databases based on ICD-9 and ICD-10 diagnostic codes; and, more recently by Metcalfe et.al for use in Read-coded (CTV3) databases.[33,37,38] We used the code lists generated from these two publications to identify the specific sets of codes that identify each of the 31 Elixhauser comorbidities in the hospital episodes and cancer registries, coded in ICD10, and in the TPP primary care data, coded in CTV3.

2.3.2 Study Population

We retrieved all clinical diagnoses in the hospital episodes, primary care (TPP only), and cancer registry datasets for events between January 1st, 2015 through December 31st, 2019. Given that only a subset of the UK Biobank participants are registered within a TPP primary care practice , we limited the hospital episodes and cancer registry diagnoses to only participants that had at least 1 clinical record in a TPP practice for any clinical finding (not limited to the 31 Elixhauser comorbidities) in the 5 years from 2015-2019. We made this decision in order to more accurately measure the prevalence of the EM comorbidities. The absence of comorbidity diagnoses in participants without primary care data linkage may be more of a reflection of certain diseases not being commonly seen in hospital records (e.g. uncomplicated diabetes) rather than truly reflecting the absence of disease. We then excluded any participants that had a mortality event in the death registry before December 31st, 2019 in order to ensure that our study sample was at risk of COVID-19 infection at the beginning of the epidemic in the UK in 2020. Next, we identified all patient diagnosis records that matched each of the 31 Elixhauser comorbidity definitions (described above). Each participant was assigned a value of 'Yes' for the

presence of a given EM comorbidity if they had at least 1 record of the clinical diagnosis in any of the three data sources (hospital episodes, primary care, or cancer registry) in the 5 years from 2015-2019. Our final analytic sample consisted of UK Biobank participants with at least 1 clinical event in the TPP primary care data from 2015-2019 (including, but not exclusive to the Elixhauser comorbidities) that were alive on January 1st, 2020. Age in years was defined as the difference between the participant's birth year (*Field #34*) and 2020. Participant's biological sex was defined as sex determined from genotyping analysis (*Field #22001*).

2.3.3 COVID-19 Severity

Severe COVID-19 infection was defined as participants that met either of the following conditions: 1) hospital inpatient diagnosis (primary or secondary) of ICD10 code U07.1 (*lab-confirmed COVID-19*) or U07.2 (*clinically/epidemiologically-diagnosed COVID-19*); or, 2) a mortality event after January 1st, 2020 with primary or contributing cause recorded with ICD-10 U07.1 or U07.2 codes. Participants that did not meet any of these conditions were assigned a value of 'No' for 'Severe' COVID-19 infection.

2.3.4 Data Analysis

Comorbidity patterns were assessed using latent class analysis (LCA) with the 31 Elixhauser comorbidities as indicators. Previous derivations of a bias-adjusted three-step LCA method required the assumption of no direct effects between covariates and the indicators used to construct the LC model. The relationship between chronic disease states (e.g. comorbidities) and age and sex has been well-documented.[39–41] We assessed the relationships between these potential confounders and each of the 31 Elixhauser comorbidity indicators using ANOVA (age) and Chi-Squared (sex) tests. Significant differences (p-value<0.001) between nearly all comorbidity indicators and both age and sex variables. In addition, we also observed significant differences (p-value <0.001) between both age and sex with respect to severe COVID-19 infection. These covariate-indicator and covariate-outcome relationships clearly violate the assumption of no direct effects between covariates and the disease indicators planned to construct our LC model. To address bias due to the direct effects of age and sex on the comorbidity

indicators and our distal outcome (severe COVID-19 infection), we followed the modified three-step approach by Vermunt et.al.[16]:

1. Estimating the LC model and the determination of the number of latent classes. Once the optimal LC model was selected, the step-one model was re-estimated to include age and sex as covariates.
2. Classifying individuals into one of the classes based on the selected step-one model, accounting for the fact that step-two classifications depend on the values of age and sex included in the step-one model.
3. Examining the relationship between classes and severe COVID-19 infection while accounting for classification errors introduced in step two. The step-three logistic regression model contained both age and sex covariates included in step-one in addition to each outcome of interest. The key modification compared to standard step-three modeling is that our classification error correction matrix was allowed to differ by age and sex.

We conducted a number of sensitivity analyses in addition to our primary analysis. First, we recalculated our outcome variable to distinguish between hospitalization (only) and mortality (with and without hospitalization). We also recalculated our outcome to distinguish between phases of the COVID-19 epidemic in the UK: events from 01-Feb-2020 to 30-Jun-2020 (Phase 1 events) and events from 01-Jul-2020 to 31-Dec-2020 (Phase 2 events). We then used a multinomial logistic regression model in Step 3 to evaluate latent class membership and our revised outcomes ('No COVID-19 Event', 'COVID-19 Hospitalization without Mortality', and 'COVID-19 Mortality'; and, 'No COVID-19 Event', 'Phase 1 event', and 'Phase 2 event'), adjusting for age and sex. Second, we stratified our sample into participants 65 years or younger (as of 2020) and participants older than 65 years; using the same Step 3 binary logistic regression model from our primary analysis to assess differences in the risk for 'severe COVID-19 infection' by age group. Finally, to investigate the potential impact of kinship between participants in our study sample, we generated two additional study samples: 1) an 'unrelated' sample by randomly breaking pairs of relatives (up to and including 3rd degree relatives, based on KING kinship coefficient

values provided by UKB (KING > 0.0442) to maintain only one member of a related pair within our study sample; and, 2) 'mixed kinship' sample by down-sampling our original study sample to the same number of participants as the 'unrelated' sample. We then compared both LCA model selection and Step 3 results for severe COVID-19 infection risk between the 'unrelated' and 'mixed kinship' samples.

In addition to the sensitivity analyses focused on study sample composition and categorization of our outcome, we conducted three analyses as illustrative examples of current techniques to compare our results against. First, we used the 'simultaneous estimate' approach, including all 31 Elixhauser comorbidities as variables in a single logistic regression model for severe COVID-19 infection and another model for COVID-19 mortality, adjusting for age and sex. We then re-ran these models with the inclusion of '# of Elixhauser comorbidities' (coded as '0', '1', and '2 or more') as an additional categorical variable. Second, we employed a 'disease-by-disease' approach, estimating separate logistic regression models for severe COVID-19 infection and COVID-19 mortality for each of the 31 Elixhauser comorbidities, adjusting for age and sex; and, applying a Bonferonni correction to the p-values to account for multiple hypothesis testing. We also re-ran these models with the '# of Elixhauser comorbidities' as an additional covariate. Finally, we assessed the count of Elixhauser comorbidities as our exposure of interest against severe COVID-19 infection in another logistic regression model, also adjusting for age and sex.

Identification of the study population, Elixhauser comorbidities, COVID-19 diagnosis, covariates, descriptive analysis, plots, figures, and sensitivity regression modelling were completed using R version 3.6.1.[42] The entire LCA modified three-step approach was completed using Latent Gold 6.0.[43]

2.4 Results

Of the 502,488 participants in the UK Biobank cohort, 176,894 participants had at least 1 diagnosis, for any condition, within the TPP linked data between 2015-2019. Of these participants, we excluded the 6,160 participants that had a mortality event in the death registry before December 31st, 2019. Our final study sample was comprised of 170,734 participants, 46.2% (n=78,814) biologically male, and on average 68.1 (SD=8.1) years of age at the end of 2019.

A total of 1,019 participants, representing 0.6% of our study sample, had a hospital inpatient COVID-19 diagnosis and/or a mortality event on or after January 1st, 2020 as of the 02 February 2021 data release. Participants with severe COVID-19 were older (72.3 years vs. 68.0 years) and more male (63.4% vs. 46.1%) compared to those without severe COVID-19 infection. Descriptive frequencies of demographic variables and the Elixhauser comorbidities by COVID-19 susceptibility and severity are provided in **Table 1**.

Approximately half of our study cohort did not have diagnoses with any of the Elixhauser comorbidities from 2015-2019 (49.4%, n=84,362). Twenty-three percent (n=40,579) of our study cohort had a diagnosis with 1 Elixhauser conditions; and, 26.8% (n=45,793) had 2 or more Elixhauser conditions. The most prevalent conditions in our cohort were uncomplicated hypertension (21.7%, n=37,074); chronic pulmonary disease (14.4%, n=24,554); uncomplicated diabetes (10.3%, n=17,627); and, obesity (9.5%, n=16,292). Prevalence of each comorbidity was higher among participants with severe COVID-19 infection for 29 of the 31 of the Elixhauser comorbidities.

2.4.1 Latent Class Analysis

To identify the optimal number of latent classes, we fit a sequence of models with 31 Elixhauser comorbidity indicators for 2 to 15 latent classes, inclusive. First, we inspected the following model parameters, tests, and fit indices: log-likelihood value (LL); Bayesian Information Criterion (BIC); sample-size adjusted BIC (SABIC); Akaike information Criterion (AIC); and, entropy R2. Lower values of BIC, SABIC, and AIC indicate better fit. Whereas, higher entropy R2 indicate better classification accuracy. These results are provided in **Figure 1** and **Table 2**. While the BIC, SABIC, and AIC continued to decrease as additional classes were added, the magnitude of decrease began to stabilize between the 5 and 7 class solutions. With the exception of the 2-class solution, the 5-class solution had the highest entropy R2 (0.6371), with slightly lower values for the 6-class (0.6062) and 7-class (0.5946) solutions; and, continued decreases in classification accuracy for solutions beyond the 7-class model.

For the 5-, 6-, and 7-class solutions, we then used item-response probabilities to assess homogeneity of the latent classes and latent class separation to select our final latent class solution. High

homogeneity (item-response probabilities that are close to 0 or 1) indicates that there is a strong relationship between the indicator (a given comorbidity in our study) and the latent class, meaning that the particular response could be determined with a high level of certainty given latent class membership.[44] A high degree of latent class separation refers to the ability to distinguish item-response probability patterns between the different classes; indicating that a response pattern describing one class describes only that class.[45] Class 1 had item-response probabilities near 0 for all comorbidity indicators across the 5-, 6-, and 7-class solutions, signifying high homogeneity for this class. Class 2 also had low item response probabilities but had less homogeneity among the indicators compared to Class 1. These solutions all two latent classes with good homogeneity: one class where 'Diabetes, uncomplicated' had an item-response probability close 1 and another latent class with 'Solid tumor, without metastasis' had an item-response probability close to 1. The 5-class solution identified a homogenous latent class where 'Hypertension, uncomplicated' had an item response probability close to 1. The additional classes generated when moving from the 5-class to the 6- and 7-class solutions also had 'Hypertension, uncomplicated' as the indicator with an item response probability close 1. These results indicated that the 6- and 7-class solutions had lower latent class separation compared to the 5-class solution. Consequently, we selected the 5-class model as our final solution based on the model statistics, within-class homogeneity, and latent-class separation.

Following selection of the 5-class model, we estimated our Step 2 model for classification including age and sex as covariates for classification as well as for covariate-indicator direct effects. Given that we observed significant differences (p-value<0.001) between nearly all comorbidity indicators and both age and sex variables, we assumed age and sex had direct effects on each disease indicator. We used the 'Step3' module in Latent Gold 6.0 to estimate the relationship between latent class membership severe COVID-19 infection with a logistic regression model using the classification posterior probabilities generated in Step 2 in addition to the inclusion of our covariates in the Step3 estimation.

2.4.2 Comorbidity and Multimorbidity Latent Classes and COVID-19

For ease of reference, we assigned each latent class a 'researcher label' (**Table 3**) based on its lead or key conditions (identified in the class-specific item response probabilities, **Figure 2**) and the median number of conditions. **Figure 3** displays the within-class prevalence as well as the difference between the within-class prevalence and overall study population prevalence for each of the 31 Elixhauser comorbidities. Odds Ratios (OR) and 95% Confidence Intervals (95% CI) for 'severe COVID-19 infection' (calculated via the bias-adjusted three-step method described previously) for each latent class compared to the reference 'Healthy' latent class are provided in **Table 4**.

The majority of our participants were assigned to the '*Healthy'* latent class. These participants were mostly female (56.3%, n=69,475); and, on average younger (67.0 years, SD=8.0) compared to the other classes. Sixty-seven percent (n=84,362) of this class had no comorbidities, 30.8% (n=38,579) had 1 comorbidity, and 2.1% (n=2,623) had 2 or more comorbidities. Participants in this class were generally healthier than the overall study population: within-class disease prevalence compared to overall study prevalence was lower for all comorbidity indicators (excluding the n=2 conditions with an overall prevalence <0.1%). Only 0.3% (n=359) of the '*Healthy*' class had severe COVID-19 infection.

The second largest class was the '*Some non-specific health conditions*' class with 17.0% (n=29,006) of the study population. Participants in this class had a median of 2 Elixhauser comorbidities, were also mostly female (52.4%, n=15,203), and had the second highest average age of the 5 latent classes (71.3 years, SD=7.1). Only 6.9% of participants in this class had a single comorbidity diagnosis (n=2,000), while 93.1% (n=27,006) had 2 or more comorbidity diagnoses. Twenty-nine of the 31 comorbidity indicators had item-response probabilities less than 20% for this class. While '*Hypertension, uncomplicated*' (49.8%) and '*Chronic pulmonary disease*' (23.2%) had item response probabilities greater than 20%, they still fell far below an 80%-100% range that would signify these conditions as lead conditions. These two comorbidities also had the highest within-class prevalence ('*Hypertension, uncomplicated*', 59.7%; '*Chronic pulmonary disease*', 29.6%) and the largest differences in within-class prevalence and overall prevalence ('*Hypertension, uncomplicated*', +38.0%; '*Chronic pulmonary*

disease', +15.2%) of the 31 comorbidity indicators. One percent of the '*Some non-specific health conditions*' class had severe COVID-19 infection (1.1%, n=320). Compared to the '*Healthy*' class, the odds of severe COVID-19 infection were 4.4 times higher for members of the '*Some non-specific health conditions*' class (OR=4.4, 95% CI: 3.4-5.7).

'*Diabetes, uncomplicated*' had an item response probability of 99.2% for our '*Diabetics with 1-2 other conditions*' latent class. Members of this class comprised 5.9% (n=10,116) of our study sample, were mostly male (57.7%, n=5,834), and ranked 3rd out of our 5 classes with respect to average participant age (70.6 years, SD=7.3%). Participants in this class had 3 median Elixhauser comorbidities, with all participants in this class having 2 or more comorbidity diagnoses. Excluding '*Diabetes, uncomplicated*', the two other comorbidity indicators with the highest within-class prevalence were '*Hypertension, uncomplicated*' (68.1%) and '*Obesity*' (30.8%). One and a half percent of this class had severe COVID-19 infection (1.5%, n=153). Compared to the '*Healthy*' class, the odds of severe COVID-19 infection were 6.4 times higher for members of the '*Diabetics with 1-2 other conditions*' class (OR=6.4, 95% CI: 4.8-8.4).

Our fourth class, the '*Cardiac multimorbidity*' class, was largely distinguished by a high item-response probability for '*Hypertension, uncomplicated*' (82.6%). Members of this class comprised 2.4% (n=4,033) of our study sample; and, had the highest proportion of males (59.4%, n=2,397), the highest age (72.3 years, SD=7.0), the largest median number of comorbidities (6) of all 5 latent classes, with all participants having 2 or more comorbidities. Excluding 'Hypertension, uncomplicated', the 3 other comorbidities with the highest difference between within-class prevalence and overall study prevalence were all cardiac-related comorbidities: '*Cardiac arrhythmia*' (+64.7%), '*Congestive heart failure*' (+51.5%), and '*Valvular disease*' (42.0%). Severe COVID-19 infection was observed among 3.6% (n=147) of this class. Members of the '*Cardiac multimorbidity*' class had the highest odds of severe COVID-19 infection of the 4 classes that were compared to the '*Healthy class*' (OR=16.6, 95% CI: 12.9-21.2).

Our final class, the '*Cancer multimorbidity*' class, was labeled for the high item-response probability for '*Solid tumor without metastasis*' (93.5%). The 'Cancer multimorbidity' class comprised 1.2% (n=2,015) of our study population, was substantially more female than our other classes (63.7%, n=691), was the 2nd youngest latent class (68.8 years, SD=8.1), had 3 median comorbidities, and all had 2 or more comorbidities. Nearly all members of this class had a 'Solid tumor without metastasis' (99.6%); with 79.7% having '*Metastatic cancer*'. Two percent of the 'Cancer multimorbidity' class had severe COVID-19 infection (n=40). Compared to the 'Healthy' class, the odds of severe COVID-19 infection were 9.9 times higher for members of the 'Cancer multimorbidity' class (OR=9.9, 95% CI: 6.8-14.3).

2.4.3 Sensitivity Analyses

Using the same 5 latent classes as our primary analysis, we conducted an additional Step 3 analysis investigating the relationship between latent class membership and 'No Event', 'COVID-19 Hospitalization without Mortality', and 'COVID-19 Mortality (with or without hospitalization)' using a multinomial logistic regression, adjusting for the same covariates as our primary analysis (age and sex). Results were consistent with our primary analysis for each of the 4 latent classes compared with the '*Healthy*' latent class: significantly increased odds for both 'COVID-19 hospitalization without mortality' and 'COVID-19 mortality (with or without hospitalization), with the strongest effect size observed for members of the '*Cardiac multimorbidity*' class (**Supplementary Table 1**). Effect sizes for this analysis compared to our primary analysis were attenuated for 'COVID-19 hospitalization without mortality'; while much stronger effect sizes, albeit less precise, were observed for 'COVID-19 mortality (with or without hospitalization)'.

After analyzing the distribution of severe COVID-19 infection events in our sample, there was a clear, bimodal distribution (**Supplementary Figure 2**). We recalculated our outcome to distinguish between these two phases of the COVID-19 epidemic in the UK: severe COVID-19 infection events from 01-Feb-2020 to 30-Jun-2020 (Phase 1 events) and events from 01-Jul-2020 to 31-Dec-2020 (Phase 2 events). Again, using the same 5 latent classes as our primary analysis, we conducted an additional Step 3 multinomial logistic regression, adjusting for age and sex, to investigate the relationship between latent

class membership and 'No Event', 'Phase 1 Event', and 'Phase 2 Event' (**Supplementary Table 2**). Results were consistent with our primary analysis: significantly increased odds of both 'Phase 1 Events' and 'Phase 2 Events' infection for the 4 latent classes compared to the 'Healthy' latent class. However, the effect sizes for the '*Some Non-Specific Health Conditions*' (OR=4.9, 95% CI: 3.3-72) and '*Diabetics with 1-2 other conditions*' classes (OR=5.0, 95% CI: 3.2-7.8) were similar for Phase 1 events, but had divergence (similar to our main findings) for Phase 2 events (OR=4.0, 95% CI: 2.8-5.7 vs. OR=7.5, 95% CI: 5.2-10.6, respectively).

We stratified our sample into participants 65 years or younger (n=63,032) and those older than 65 years (107,702) to investigate potential differences in latent class model selection or latent class membership and severe COVID-19 infection. The criteria we used to select the 5-class model in our primary analysis held for both the '*65 Years or Younger*' and '*Older than 65 Years*' samples (model fit statistics, latent class heterogeneity, and latent class separation), indicating that the 5-class solution was optimal for both samples. All model fit statistics and item-response probabilities for these two samples are provided in **Supplementary Figures 3-6**. Results from the Step 2 and Step 3 LCA process to assess the relationship between latent class membership and severe COVID-19 infection were consistent with our main findings for the '*Older than 65 Years*' samples (**Supplementary Tables 3 and 4**). However, among the *'65 years or younger*' sample, members of the '*Some Non-Specific Health Conditions*' class lower odds of severe COVID-19 infection than our primary sample results (OR=2.8, 95% CI: 1.3-5.7 vs. OR=4.39, 95% CI: 3.36-5.74). In contrast, members of the '*Diabetics with 1-2 other conditions*' class compared to the '*Healthy*' class had higher odds of severe COVID-19 infection than our primary sample results (OR=7.6, 95% CI: 4.4-12.9 vs. OR=6.36, 95% CI: 4.82-8.40).

To investigate the potential impact of kinship between participants in our study, we generated two additional samples: 1) an '*unrelated*' sample (n=151,623) by randomly breaking pairs of relatives (up to and including 3rd degree relatives, based on KING kinship coefficient values > 0.0442, to maintain only one member of a related pair within our study sample; and, 2) '*mixed kinship*' sample by down-sampling our original study sample to the same number of participants as the 'unrelated' sample by keeping all

unrelated participants and adding a random selection of participants with 3rd degree relatives to reach the same sample size as our 'unrelated' sample (n=151,623). All model fit statistics and item-response probabilities for the two samples are provided in **Supplementary Figures 7-10**. The rationale used to select the 5-class model from our main results held for both the '*unrelated*' and '*mixed kinship*' samples: similar model fit statistics, latent class heterogeneity, and latent class separation. We then followed the same Step 2 and Step 3 LCA process to assess the relationship between latent class membership and severe COVID-19 infection in both samples. Results for severe COVID-19 infection were consistent between our primary study sample as well as the '*unrelated*' and '*mixed kinship*' samples (**Supplementary Table 5**).

Regression results for our '*simultaneous*' and '*disease-by-disease*' examples (both with and without the additional *'# of Elixhauser comorbidities*' covariate) are provided in **Supplementary Tables 6 and 7**. Our '*simultaneous estimate*' illustrative example identified significantly higher odds of severe COVID-19 infection for 16 Elixhauser comorbidities after adjusting for age and sex as covariates. Only 13 of these comorbidities remained significant after adding *'# of Elixhauser comorbidities*' as a covariate to the model. For our '*disease-by-disease*' estimate, the odds of severe COVID-19 infection were higher for 27 of the Elixhauser comorbidities, adjusting for age and sex and correcting p-values for multiple hypothesis testing. Only 12 of these comorbidities remained significant after the addition of the *'# of Elixhauser comorbidities*' covariate. Generally, the strength of the effects was much higher for the '*disease-by-disease*' estimates than the '*simultaneous estimate*' results. Finally, for our logistic regression model treating the '# of Elixhauser comorbidities' as our exposure, the odds of severe COVID-19 infection increased by 1.39 for each additional comorbidity (OR=1.39, 1.36,1.42), adjusting for age and sex.

2.5 Discussion

We identified 5 distinct comorbidity patterns from 31 disease indicators, assessed using clinical diagnosis records from UK Biobank's comprehensive EHR data linkage between 2015-2019. Our results identified significantly increased risk for severe COVID-19 infection for our '*Some Non-Specific Health Conditions*', '*Diabetics with 1-2 other conditions*', '*Cardiac multimorbidity*', and '*Cancer multimorbidity*' latent classes compared to our '*Healthy*' latent class. In addition, our results identified substantial heterogeneity in the effect sizes of severe COVID-19 infection risk between our comorbidity latent classes.

We demonstrated that not all combinations of comorbid diseases have equal importance with respect to describing risk for severe COVID-19 infection. Our '*Diabetics with 1-2 other conditions'* and '*Cancer multimorbidity*' had the same median # of Elixhauser comorbidities (3) but had substantial variation in the strength of their relationship to severe COVID-19 infection. The '*Cancer multimorbidity*' class had a higher odds of severe COVID-19 infection (OR=9.9) compared to the '*Diabetics with 1-2 other conditions*' (OR=6.4). Other studies have reported associations between 'counts' of comorbidities and COVID-19 outcomes.[46–49] As a illustrative example, we examined the association between a simple count of Elixhauser comorbidities as our exposure and severe COVID-19 infection as our outcome, showing significant increases in risk for severe COVID-19 infection for each additional comorbid condition. However, the results of our main analysis illuminate the vulnerabilities of using these simple counts: 1) describable combinations of comorbidities exist in real-world patient populations; and, 2) that the components of these comorbidity combinations impact risk for severe COVID-19 infection. Consequently, our findings stress the importance of considering heterogeneity when dealing comorbidities; and, the need research to consider complex patterns of disease and not assume that comorbidity counts are sufficient for confounding control.

Our findings also highlight the need to consider the presence of multiple comorbidities when studying COVID-19 infection severity. More than a quarter of our sample (26.8%, n=45,793) had 2 or more comorbidities. In another study of primary care records for 17 million adults in England,

Williamson et.al. identified 12 health conditions that had a significantly increased risk for COVID-19-related death; and, 1 health condition (hypertension or high blood pressure) that was associated with decreased risk for COVID-19 mortality.[50] Our 'simultaneous estimate' approach (a crude approximation of the methods by Williamson et.al.) found consistent findings (increased risk of COVID-19 mortality) for diabetes, kidney disease, chronic liver disease, and other neurological disease; while we found increased COVID-19 mortality risk for hypertension compared to the decreased risk reported by Williamson et.al. However, when adding '# of comorbidities' to our 'simultaneous estimate' model, our hypertension association with mortality became null, while all other findings remained significant. A recent comment by Westreich et.al. identified that these findings are particularly vulnerable to the 'Table 2 Fallacy' given that all hazard ratios were reported from a single regression model.[51] In addition, Williamson et.al. do not present any description of the presence of multiple comorbid diseases in their sample; consequently, in contrast to our latent class results, it is unclear how well their model controlled for interactions among many related disease states among patients with more than one disease. Our 'disease-by-disease' estimate (an crude, alternative approach to Williamson et.al.) identified significantly increased odds of COVID-19 mortality for 25 of the 31 Elixhauser comorbidities, replicating Williamson et.al.'s findings for diabetes, kidney disease, chronic liver disease, and other neurological disease, although with much larger effect sizes. When adding '# of comorbidities' as an additional covariate in the 'disease-by-disease' estimation models, only chronic liver disease, kidney disease, and other neurological disease remained significant. This suggests that the 'disease-by-disease' approach may be very sensitive to the presence of multiple comorbidities. However, neither of these approaches is able to identify specific combinations of comorbidities that are important to COVID-19 infection severity. In contrast to these common, current research practices, our LCA method was able to not only identify comorbidity patterns, but also was able to capture differences in the relationship between specific comorbidity patterns and severe COVID-19 infection.

Our '*Diabetics with 1-2 other conditions*' class had higher odds of severe COVID-19 infection when compared to the 'Healthy' class (OR=6.4) than the comparison of 'Some Non-Specific Health

Conditions' to the 'Healthy' class (OR=4.4). This additional risk may be partially explained by evidence asserting that diabetes is implicated in COVID-19 severity. A meta-analysis of 33 studies (16,003 patients) found that diabetes in patients with COVID-19 was associated with a two-fold increase in mortality.[52] In a study of 2,433 COVID-19 patients in China, *Wang et.al.* found that patients with elevated blood glucose levels 3.22 times more likely to die of COVID-19.[53] Differences in blood glucose levels, reflecting differences in diabetic control, may provide a potential explanation for the discrepancy in COVID-19 severity between the two diabetes classes. *Li et.al.* found that COVID-19 patients with newly diagnosed diabetes had the highest risk of all-cause mortality compared with COVID-19 patients with known diabetes, hyperglycemia, and normal glucose.[54] The important clinical consideration for our study is that diabetes, especially in the presence of additional comorbidities, may be an important driver of increased risk for severe COVID-19 infection.

Our '*Cardiac multimorbidity*' class had by far the highest odds of severe COVID-19 infection of the four latent classes that were compared to the 'Healthy' reference class (OR=16.6). While heart-related comorbidities featured prominently in this class, the median # of comorbidities in this class (6) was double that of the '*Diabetics with 1-2 other conditions*' (3) class and the '*Cancer multimorbidity*' (3) class; and, triple that of the '*Some Non-Specific Health Conditions*' (2) class. The mean age of the '*Cardiac multimorbidity*' class (72.3 years) was similar to that of the '*Some Non-Specific Health conditions*' (71.3 years), which in addition to our methods for controlling confounding by age, suggests that the most notable difference potentially driving risk for severe COVID-19 infection between these classes is with respect to the # of comorbidities. Consequently, while there are many studies indicating that heart-related comorbidities increase risk for severe COVID-19, our study cannot differentiate between a 'heart disease specific' effect and the effect of overall very poor health (the presence of many comorbidities).[55–57] Consequently, our findings support current COVID-19 efforts targeting individuals in very poor health for vaccination or other interventions that may prevent COVID-19 infection or reduce severity of COVID-19 infection.

Our '*Cancer multimorbidity*' class also had substantially high odds of severe COVID-19 infection compared to the 'Healthy' reference class. In addition, the effect size was much higher for the '*Cancer multimorbidity*' class (OR=9.9) than the '*Diabetics with 1-2 other conditions*' class (OR=6.4), despite having the same median # of comorbidities (3). Current research on pre-existing cancer diagnoses and COVID-19 severity is mixed. Krasnow et.al. did not identify any association between severe COVID-19 illness and cancer in a cohort of mostly African American patients in the United States.[58] A retrospective study of hospitalized patients in Wuhan, China identified that cancer patients had a higher risk of mortality than noncancer patients, after applying propensity score matching.[59] While another study from China showed that patients with cancer had poorer COVID-19 outcomes, a more recent response to this article asserted that 'current evidence remains insufficient to explain a conclusive association between cancer and COVID-19'.[60,61] Our research indicates that the '*Cancer multimorbidity*' class has an important relationship to severe COVID-19 infection; however, the current mixed evidence available around cancer and COVID-19 indicates that more targeted research beyond our work is required to understand this relationship.

We observed similar results when re-coding our outcome to distinguish between COVID-19 hospitalization and mortality, albeit with larger, less precise effect sizes for mortality (likely due to the small number of mortality events in the study sample). When investigating differences in severe COVID-19 infection by phase of the pandemic in the UK, effect sizes for the '*Some Non-Specific Health Conditions*' and '*Diabetics with 1-2 other conditions*' were similar in Phase 1 (OR=4.9 vs. OR=5.0), but had divergence (similar to our main findings) for Phase 2 events (OR=4.0 vs. OR=7.5). The B.1.1.7 COVID-19 variant was first identified from patients with COVID-19 in the south east of England in early October 2020.[62] Recent studies have shown that not only is the B.1.1.7 variant associated with increased transmission, it is also associated with increased risk for severe outcomes, including mortality.[62–65] A recent study found that the hazard for death for patients with the B.1.1.7 variant compared to patients without the B.1.1.7 variant was significantly higher for patients with one or two or more comorbidities.[62] Given these findings, it is plausible that the divergence observed during Phase 2 could be at least partially

attributed to additional risk from the B.1.1.7 variant for individuals with existing comorbidities; or, specifically for individuals with diabetes and diabetes-related comorbidities. However, we did not identify any other studies that specifically investigated whether the B.1.1.7 variant is more severe among either individuals with multiple comorbidities or for individuals with specific comorbidities (e.g. diabetes).

We conducted two additional sensitivity analyses beyond what has already been described above. In our stratified analysis by age, our results for the 'Older than 65 years' sample were consistent with our primary analysis. However, among the '65 years or younger' sample, we observed a smaller effect size for our '*Some Non-Specific Health Conditions*' class compared to our primary analysis (OR=2.8 vs. OR=4.4) and a larger effect size for our '*Diabetics with 1-2 other conditions*' class compared to our primary sample results (OR=7.6 vs. OR=6.4). It is unclear what may be driving this difference given that the # of comorbidities and composition of comorbidities is nearly the same for both the '65 years or younger' and the primary study sample. We investigated the potential impact of kinship between participants in our study by generating two additional samples and comparing our latent class selection and Step 3 logistic regression results. We found no indication that kinship had any impact on our study findings, suggesting that kinship may be more important for genetic studies rather than purely observational studies in the UK Biobank.

There are a few important limitations of our research. We did not measure causal relationships between comorbidity latent classes and our COVID-19 outcomes; our research was intended to identify specific comorbidity patterns that may be important for future causal-inference based research. Consequently, our findings should be interpreted as evidence of variation in COVID-19 severity across distinct comorbidity patterns. Cases identified via primary care data were ascertained only from TPP EHR data. The absence of a published, validated code list for the Elixhauser Comorbidities in SNOMED-CT prevented our inclusion of cases identified in EMIS practices in the past 5 years. Our definition of Elixhauser Comorbidities was derived entirely from diagnosis codes. The inclusion of additional data points (e.g. prescriptions and/or laboratory values) could have not only increased our case numbers for

some diseases, but also could have improved specificity and sensitivity of comorbidity diagnosis. There is significant potential for within-disease heterogeneity due to coding practices by different health care providers; as well as vulnerabilities due to provider ascertainment of sufficient disease to warrant recording a diagnosis code. In addition, diagnosis codes for COVID-19 infection among hospital inpatients or mortality registry records did not allow us to differentiate between COVID-19 variants that emerged later in 2020 and have known implications on COVID-19 infection severity.[66]

To our knowledge, this is the first application of our specific LCA estimation with a distal outcome that used diseases as indicators. Our LCA method is an improvement over previous derivations of the bias-adjusted three-step method as it did not require the assumption of no direct effects between covariates and the indicators used to construct the LC model. This was particularly important in our LCA given the well-documented impacts of age and sex on disease states. While LCA has been used in other contexts to describe comorbidity patterns, it has not been used to study comorbidity patterns in the context of COVID-19 outcomes. Our use of LCA enabled unique insight into the specific comorbidity patterns that relevant to COVID-19 outcomes. Our 'comorbidity patterns' have more relevance to real-world settings given the large prevalence of comorbidity and multiple comorbidities, particularly among older adults.

There are a number of other strengths of our research. COVID-19 studies that leverage UK Biobank data frequently use self-reported conditions, biomarkers, and other variables that were measured during the Baseline Assessment visit between 2006 and 2010. Each of our conditions was measured via diagnosis codes, which required a healthcare provider to record the diagnosis based on a clinical evaluation. Furthermore, we did not need to translate any of our diagnosis code lists across medical vocabularies (e.g. from CTV-3 to SNOMED-CT; or ICD-10 to CTV3), potentially losing or modifying the original clinical meaning of the diagnosis. Our code lists for the Elixhauser comorbidities were developed specifically for each medical vocabulary. We also did not select specific comorbidities for investigation based on our own opinions, information from prior research, or with the intention to validate a specific hypothesis. Rather, the use of a common measure of comorbidities used across many research

settings ensured that the comorbidities we included for investigation were comprehensive. Finally, our study was able to leverage the UK Biobank's primary care data linkage for COVID-19 investigation. This strengthened our ability to completely ascertain of health status of UK Biobank participants seen within TPP GP practices.

As the research community continues to evolve to meet the challenges of the COVID-19 epidemic worldwide, our research illustrates the complexity of comorbidities, particularly among older adults. As data continues to accumulate, it is tempting to investigate many potential 'risk factors' simultaneously in a single regression model. While Epidemiologists have stressed the importance of avoiding this practice due to the 'Table 2 Fallacy', our research also highlights another major vulnerability to this approach: chronic disease rarely occurs in isolation and specific interrelated comorbidities may more accurately describe risk compared to singular analysis or simple comorbidity counts. We hope that these findings will signal other researchers to carefully consider the impact of individual and multiple comorbidities in the context of their COVID-19 research.

All UK Biobank data was accessed in accordance with GlaxoSmithKline's UK Biobank Application #20361.

Chapter 2, in full, is currently being prepared for submission for publication of the material. Davitte, J.; Pines, H.; Martin, N.; Salem, R.; Brodine, S.; Shaffer, R. The dissertation author was the primary investigator and author of this material.

Table 1: Characteristics of the Study Population, Overall and by COVID-19 Severity, UK Biobank

Characteristic	Overall (N=170,734)	Severe COVID-19 Infection[1] Yes (N=1,019)	Severe COVID-19 Infection[1] No (N=169,715)
Age (Years)			
Mean (SD)	68.1 (8.1)	72.3 (7.5)	68.0 (8.0)
Range	49.0 - 84.0	51.0 - 83.0	49.0 - 84.0
Sex, Male, n (%)	78814 (46.2%)	646 (63.4%)	78168 (46.1%)
# of Elixhauser Conditions from 2015-2019			
Mean (SD)	1.1 (1.6)	3.0 (2.6)	1.1 (1.5)
Range	0.0 - 17.0	0.0 - 13.0	0.0 - 17.0
Diagnosis with Elixhauser Comorbidity, 2015-2019, n (%)			
Alcohol abuse	3026 (1.8%)	55 (5.4%)	2971 (1.8%)
Anemia, deficiency	4157 (2.4%)	78 (7.7%)	4079 (2.4%)
Anemia, blood loss	74 (0.0%)	1 (0.1%)	73 (0.0%)
Cardiac arrhythmia	13229 (7.7%)	222 (21.8%)	13007 (7.7%)
Congestive heart failure	3472 (2.0%)	85 (8.3%)	3387 (2.0%)
Coagulopathy	1002 (0.6%)	20 (2.0%)	982 (0.6%)
Depression	7283 (4.3%)	113 (11.1%)	7170 (4.2%)
Diabetes, complicated	2642 (1.5%)	62 (6.1%)	2580 (1.5%)
Diabetes, uncomplicated	17627 (10.3%)	279 (27.4%)	17348 (10.2%)
Drug abuse	188 (0.1%)	3 (0.3%)	185 (0.1%)
Fluid/electrolyte disorders	4168 (2.4%)	140 (13.7%)	4028 (2.4%)
HIV/AIDS	31 (0.0%)	0 (0.0%)	31 (0.0%)
Hypertension, complicated	121 (0.1%)	7 (0.7%)	114 (0.1%)
Hypertension, uncomplicated	37074 (21.7%)	508 (49.9%)	36566 (21.5%)
Hypothyroidism	7347 (4.3%)	83 (8.1%)	7264 (4.3%)
Liver disease	2465 (1.4%)	62 (6.1%)	2403 (1.4%)
Lymphoma	1006 (0.6%)	21 (2.1%)	985 (0.6%)
Metastatic cancer	1859 (1.1%)	43 (4.2%)	1816 (1.1%)
Other neurological disorders	6034 (3.5%)	168 (16.5%)	5866 (3.5%)
Obesity	16292 (9.5%)	156 (15.3%)	16136 (9.5%)
Paralysis	840 (0.5%)	36 (3.5%)	804 (0.5%)
Peptic ulcer disease	1545 (0.9%)	19 (1.9%)	1526 (0.9%)
Peripheral vascular disease	3056 (1.8%)	68 (6.7%)	2988 (1.8%)
Psychoses	710 (0.4%)	13 (1.3%)	697 (0.4%)
Chronic pulmonary disease	24554 (14.4%)	286 (28.1%)	24268 (14.3%)
Pulmonary circulation disorder	1240 (0.7%)	24 (2.4%)	1216 (0.7%)
Rheumatoid arthritis	4801 (2.8%)	73 (7.2%)	4728 (2.8%)
Renal failure	4438 (2.6%)	125 (12.3%)	4313 (2.5%)
Solid tumor without metastasis	10034 (5.9%)	148 (14.5%)	9886 (5.8%)
Valvular disease	3849 (2.3%)	80 (7.9%)	3769 (2.2%)
Weight loss	2208 (1.3%)	38 (3.7%)	2170 (1.3%)

[1]*Severe COVID-19 infection defined as: 1) hospital inpatient diagnosis U07.1 or U07.2; or, 2) a mortality event after January 1st, 2020 with U07.1 or U07.2 codes.*

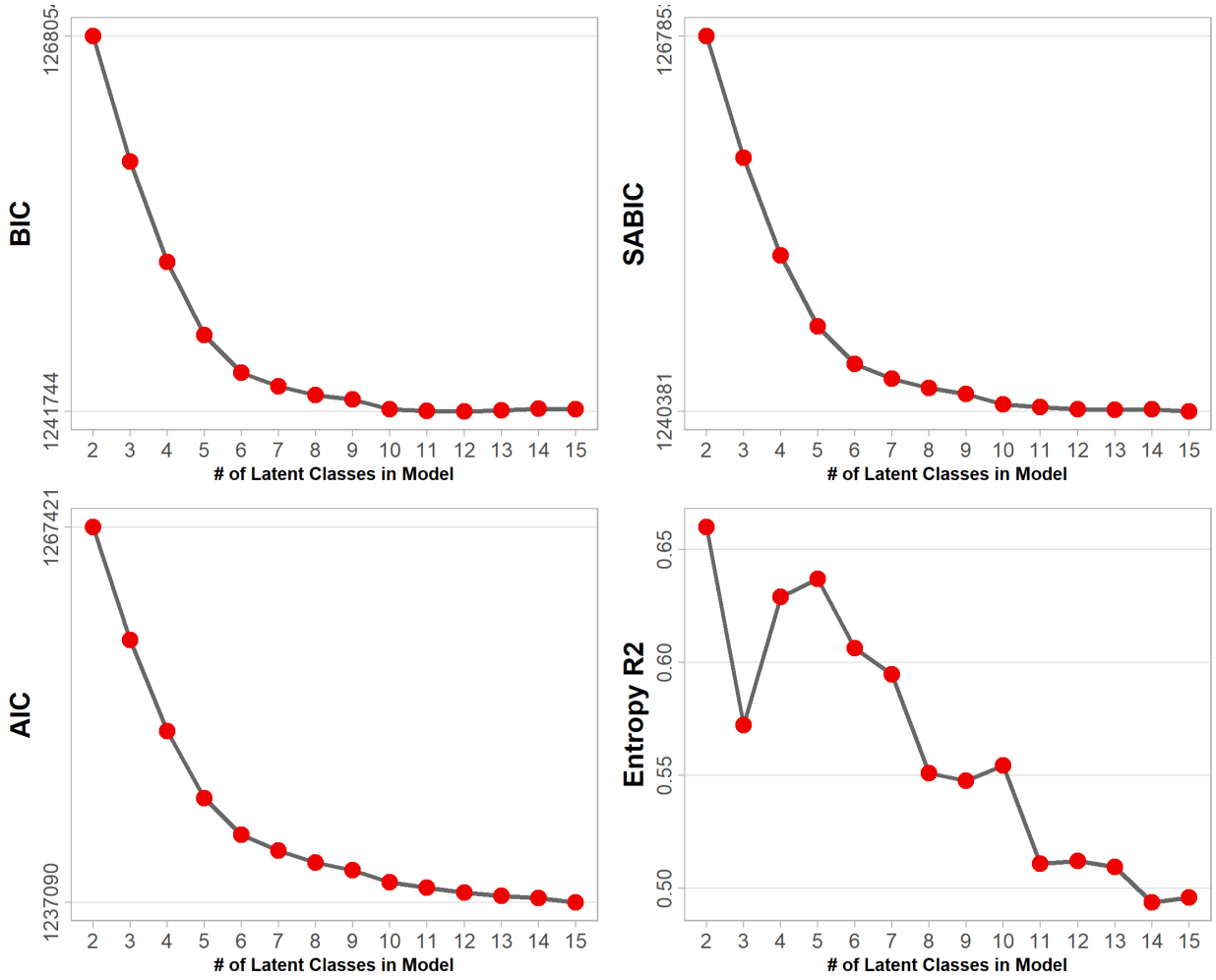


Figure 1: Latent Class Analysis Model Results for 2-15 Class Solutions

Table 2: Latent Class Analysis Model Results for 2-15 Class Solutions

Model	LL	BIC(LL)	SABIC (LL)	AIC(LL)	Entropy R^2
2-Class	-633647.2887	1268053.593	1267420.578	1267853.376	0.66
3-Class	-629049.4407	1259243.428	1258288.882	1258941.514	0.5723
4-Class	-625333.8182	1252197.715	1250921.636	1251794.104	0.6289
5-Class	-622596.2246	1247108.059	1245510.449	1246602.751	0.6371
6-Class	-621074.5739	1244450.29	1242531.148	1243843.283	0.6062
7-Class	-620401.1449	1243488.963	1241248.29	1242780.26	0.5946
8-Class	-619905.7985	1242883.802	1240321.597	1242073.401	0.551
9-Class	-619551.3064	1242560.349	1239676.613	1241648.251	0.5477
10-Class	-619032.3539	1241907.976	1238702.708	1240894.18	0.5543
11-Class	-618780.4978	1241789.795	1238262.996	1240674.303	0.5109
12-Class	-618564.972	1241744.275	1237895.944	1240527.085	0.5121
13-Class	-618398.7832	1241797.429	1237627.567	1240478.542	0.5095
14-Class	-618277.7868	1241940.968	1237449.574	1240520.383	0.4938
15-Class	-618066.2187	1241903.363	1237090.437	1240381.081	0.4959

Table 3: Features of Researcher Labeled Latent Classes for the 5-Class LCA Model

Latent Class	**Researcher Label**	**# of Participants**	**# of Elixhauser Comorbidities** *Median (Q1,Q3)*	**Age, Years** *Mean (SD)*	**Male** *N (%)*
1	*"Healthy"*	125,564 (73.5)	0 (0,1)	67.0 (8.0)	56,089 (44.7)
2	*"Some Non-Specific Health Conditions"*	29,006 (17.0)	2 (2,3)	71.3 (7.1)	13,803 (47.6)
3	*"Diabetics with 1-2 other conditions"*	10,116 (5.9)	3 (2,4)	70.6 (7.3)	5,834 (57.7)
4	*"Cardiac multimorbidity"*	4,033 (2.4)	6 (5,8)	72.3 (7.0)	2,397 (59.4)
5	*"Cancer multimorbidity"*	2,015 (1.2)	3 (2,5)	68.8 (8.1)	691 (34.3)

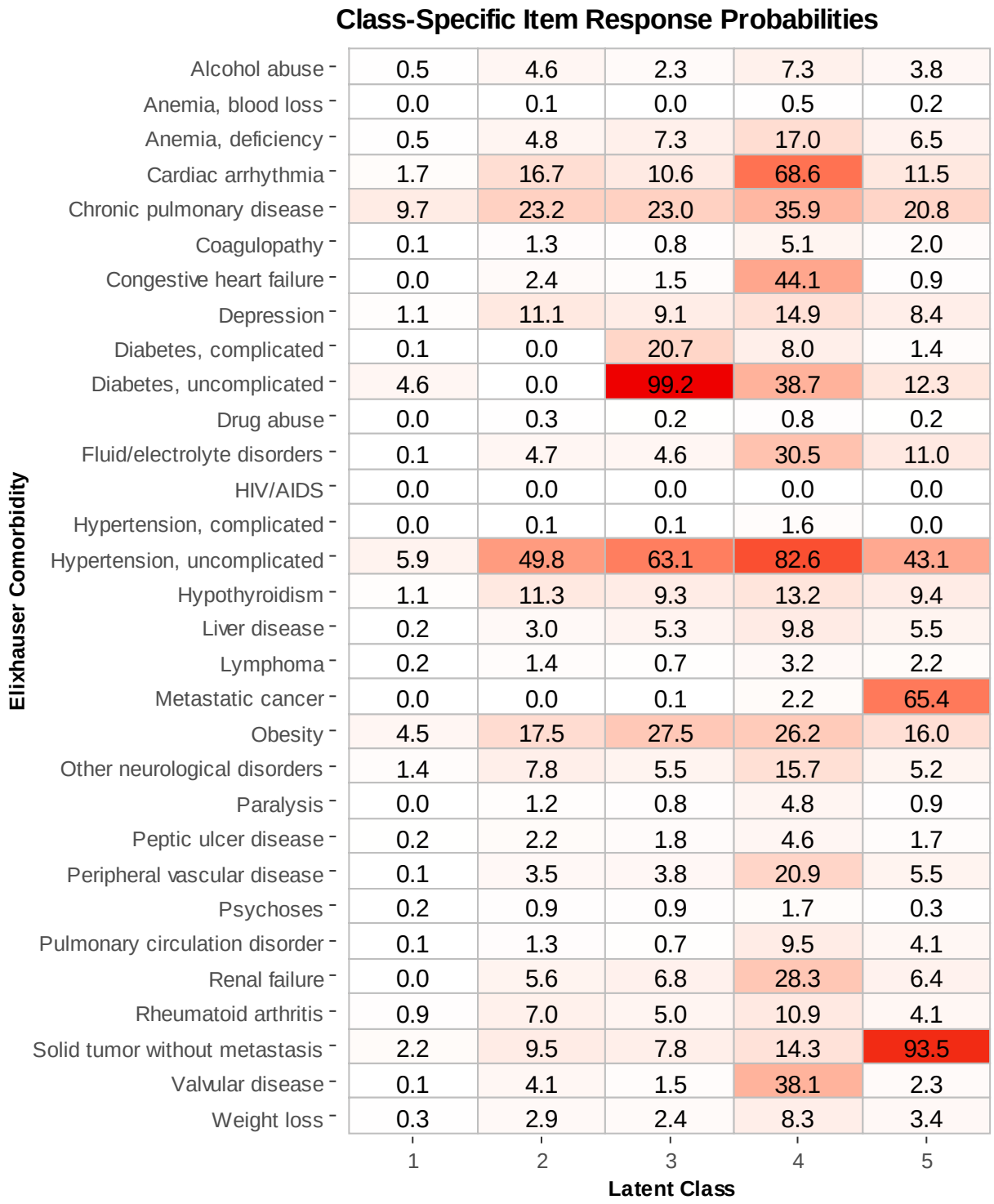


Figure 2: Class-Specific Item Response Probabilities for 31 Elixhauser Comorbidity Indicators in 5-Class Latent Class Solution.

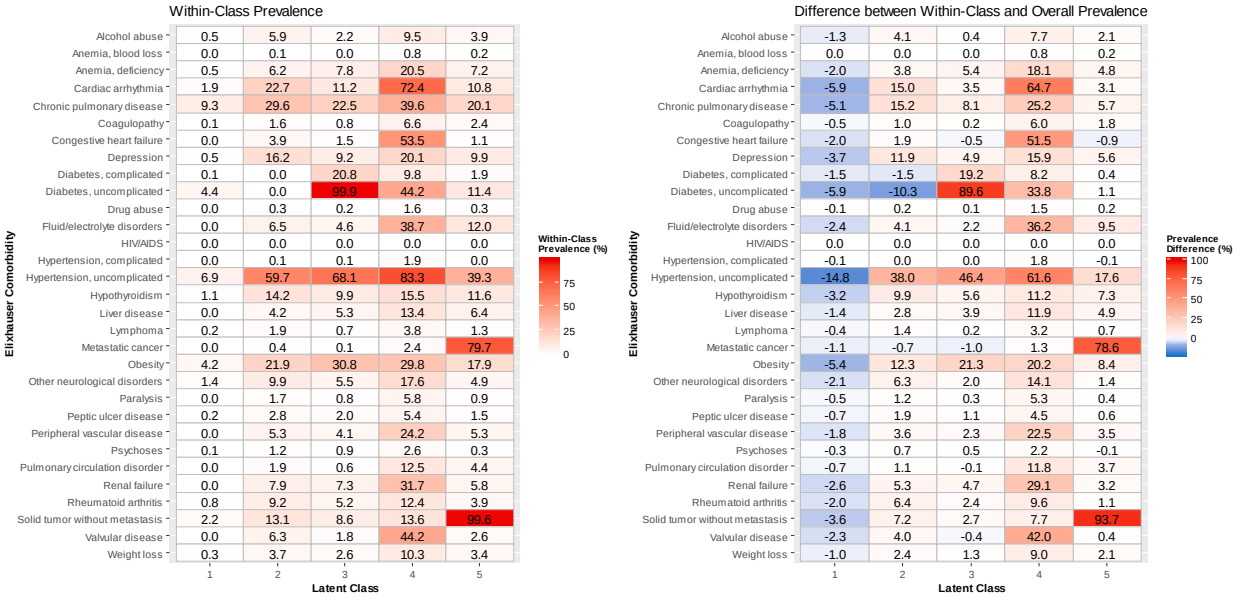


Figure 3: Within-Class Prevalence of Elixhauser Comorbidities and Difference Between Within-Class Prevalence and Overall Sample Prevalence by Latent Class

Table 4: Odds Ratio (OR) and 95% Confidence Interval (95% CI) of severe COVID-19 infection by comorbidity Latent Class membership, adjusting for participant age and sex.

Latent Class	**Researcher Label**	**# of Participants** *N (%)*	**Severe COVID-19 Infection** *N (%)*	**OR (95% CI)**
1	*"Healthy"*	125,564 (73.5)	359 (0.3)	Reference
2	*"Some Non-Specific Health Conditions"*	29,006 (17.0)	320 (1.1)	4.4 (3.4-5.7)
3	*"Diabetics with 1-2 other conditions"*	10,116 (5.9)	153 (1.5)	6.4 (4.8-8.4)
4	*"Cardiac multimorbidity"*	4,033 (2.4)	147 (3.6)	16.6 (12.9-21.2)
5	*"Cancer multimorbidity"*	2,015 (1.2)	40 (2.0)	9.9 (6.8-14.3)

Chapter 3: Polygenic Risk for Severe COVID-19 Infection and Risk for 31 EHR-Derived Comorbidities in the UK Biobank Cohort

3.1 Abstract

Polygenic risk scores (PRS) provide an immediate opportunity to assess whether the same genetic composition that increases an individual's risk for severe COVID-19 infection is shared with other important diseases. We investigated the associations between severity to COVID-19 infection genetic liability (defined via COVID-19 severity PRSs) and 31 comorbidity indicators assessed using linked electronic health record (EHR) data over the past 20 years. We developed our PRSs using the COVID-19 Host Genetics Initiative's GWAS of 'very severe respiratory confirmed COVID-19' (the largest COVID-19 GWAS to date); and, used UK Biobank's extensive linked EHR data to evaluate the relationship between the PRSs and comorbidity diagnoses. We constructed 31 case/control samples of unrelated, European ancestry participants for each of the Elixhauser comorbidities from 502,493 UK Biobank participants. We constructed PRSs across a specific range of p-value thresholds for 28 of the 31 case/control samples; and, used the PRS principal component analysis (PRS-PCA) approach to evaluate the relationship between the severe COVID-19 infection PRS and the comorbidity outcome. Our research indicates that the same genetic composition that increases an individual's risk for COVID-19 may also influence their risk for diabetes, hypertension, obesity, and renal failure.

3.2 Introduction

As of February 2021, the severe acute respiratory syndrome coronavirus 2 (SARS-COV-2 / COVID-19) has now infected more than 109 million individuals globally, resulting in 2.4 million deaths.[31] The United Kingdom currently has had more than 4 million confirmed cases and currently has the highest number of COVID-19 deaths per 100,000 population worldwide (176.90 deaths / 100,000 population).[32]

In response to this unprecedented pandemic, large genetic studies, including United Kingdom's UK Biobank[5] study, are actively working on identifying specific genetic variants that increase an individual's risk for severe COVID-19 infection.[18,67,68] Meta-analysis of genome-wide association study

(GWAS) results from 34 studies across 16 countries as part of the COVID-19 Host Genetics Initiative (HGI) found seven genomic regions associated with severe COVID-19 infection, harboring genes which regulate immune function or play a role in lung diseases.[69]

The specific genetic variants associated with COVID-19 severity identified through analyses such as the COVID-19 HGI meta-analyses may represent important opportunities for the development of novel drug targets; however, typically the magnitude of effect and variance explained by a single genetic variant is small and individually have limited utility when evaluating genetic liability for a given trait and/or health outcome(s).[18–20] The 'Common Disease, Common Variant' hypothesis posits that genetic variants with appreciable frequency in the population at large, but relatively low probability that variant carriers will express the disease (penetrance), are the major contributors to genetic susceptibility to common diseases.[21] The cumulative risk derived from aggregating contributions of the many common and uncommon variants associated with a complex trait or disease, such as COVID-19, is referred to as a polygenic risk score (PRS).[22] PRSs are commonly defined as the sum of trait genome-wide-associated (single nucleotide polymorphisms (SNPs) weighted by their effect sizes to provide an overall measure of an individual's genetic liability to that trait or disease.[23] Consequently, PRSs can achieve substantially greater predictive power for a given trait by including a larger number of SNPs in the PRS compared to restricting to only SNPs that reach GWAS genome-wide significance (e.g. $p < 5x10^{-8}$).[24] While PRSs have many applications, they have been used extensively to identify shared genetic etiology between two traits.[25,25–29] For example, Andersen et.al. found that common polygenic risk contributes to susceptibility to both major depressive disorder and alcohol dependence.[26] Another study of Parkinson's disease and blood levels of 370 lipid species found shared genetic etiology between Parkinson's disease and 25 lipids.[25]

Given the emergence of COVID-19 in the end of 2019, traditional observational research methods are unable to investigate the impact of COVID-19 on the development of new diseases and/or the exacerbation of existing comorbidities well beyond 1-year post-infection. Although the respiratory system is the primary site affected by the COVID-19 virus, infection has also been proven to be a major

threat to other organ systems, including cardiovascular, gastrointestinal, renal, central nervous, and reproductive systems.[17] Consequently, there is potential for long-term impact of COVID-19 on not only respiratory diseases, but potentially across the disease-phenome.

In this work, we investigated the associations between genetic liability to severe COVID-19 infection and 31 comorbidity phenotypes derived from linked electronic health record (EHR) data in the past 20 years. We hypothesized that increased polygenic risk for severe COVID-19 infection is positively associated with risk for some of these 31 comorbidities. In identifying comorbid diseases with shared genetic risk to COVID-19 severity, our work aims to inform hypotheses for specific genetic causal inference efforts; and, identify health outcomes that warrant additional scrutiny once sufficient longitudinal data has accumulated.

3.3 Materials and Methods

3.3.1 Target Data: UK Biobank

We used data from the UK Biobank, a prospective cohort study providing detailed characterization of over half a million UK-based persons aged 40-69 years at recruitment from 2006 to 2010, with continuous follow-up to present day through additional, bespoke data collection efforts as well as regular linkage to National Health Service (NHS) electronic health record (EHR) data and other registries (e.g. Cancer and Mortality).[5] Participants (n=502,493) were recruited at 22 centers throughout the UK, providing socioeconomic and ethnic heterogeneity as well as urban-rural mix. Data collection at the baseline assessment visit included: electronic signed consent; a self-completed touch-screen questionnaire; brief computer-assisted interview; physical and functional measures; and collection of blood, urine, and saliva. In response to the COVID-19 epidemic, UK Biobank is providing regular releases of diagnostic COVID-19 testing data, GP (primary care) data provided directly by the system suppliers, hospital inpatient data, critical care data, and mortality data to facilitate research into the determinants and consequences of COVID-19.[6]

Our study leveraged the following data sources: baseline assessment data for demographics; hospital episodes for inpatient clinical diagnoses, released on 22 February 2021, coded using the

International Classification of Diseases, Tenth Revision (ICD-10); mortality events from the death registry, released on 16 February 2021, coded in ICD-10; cancer diagnoses, released in March 2019, coded in ICD-10; primary care data supplied by 'The Phoenix Partnership' (TPP) system provider for clinical diagnoses, released 14 April 2021, coded using the Clinical Terms Version 3 (CTV3). At the time of this research, there was no validated, published definition of the Elixhauser comorbidities using SNOMED-CT terms in the UK. Consequently, we did not leverage primary care data supplied by the 'Egton Medical Information Systems' (EMIS) system provider given that clinical diagnoses in from 2015-2019 were largely coded in SNOMED-CT terms.

Comorbidity summary measures have been developed to help classify patients according to their overall disease burden. Elixhauser et.al defined a set of 31 comorbidity indicators, which have been translated by Quan et.al. for use in administrative databases based on ICD-9 and ICD-10 diagnostic codes; and, more recently by Metcalfe et.al for use in Read-coded (CTV3) databases.[33,37,38] We used the code lists generated from these two publications to identify the specific sets of codes that identify each of the 31 Elixhauser comorbidities in the hospital episodes, mortality, and cancer registries, coded in ICD10, and in the TPP primary care data, coded in CTV3.

A detailed flow diagram for our case/control selection algorithm that was repeated for each of the 31 Elixhauser comorbidity samples is provided in **Figure 4**. First, we restricted UK Biobank participants to only those of genetic 'European ancestry' (Field #21000). We then retrieved all clinical diagnoses in the hospital episodes, primary care (TPP only), death registry, and cancer registry datasets for events between January 1st, 2000 through December 31st, 2020. We identified all patient diagnosis records that matched each of the 31 Elixhauser comorbidity definitions. We used the following case and control selection methodology for each of the 31 Elixhauser comorbidity samples. Individuals that had 1 or more diagnoses for a given comorbidity were defined as an 'eligible case'. For individuals that were not 'eligible cases', we required 'eligible controls' to have 1 or more diagnosis (for any reason, not exclusive to the Elixhauser comorbidities) in the TPP data from 2000-2019. We imposed this requirement to ensure that we had complete observation of these individuals across all our included data sources, ensuring that

'eligible controls' could not have had an Elixhauser comorbidity diagnosis (e.g. met our case definition) in a data source in which their records were not linked for our study. We then assessed whether 'eligible cases' had a 1st, 2nd, or 3rd degree relative (KING > 0.0442) among either 'eligible cases' or 'eligible controls' using the KING kinship coefficient values provided by UKB. If the 'eligible case' had no relatives among 'eligible cases' or 'eligible controls', the participant was selected as a 'case'. If another case was a relative, we randomly selected 1 member of the 'eligible case' pair as a 'case'. Among 'eligible controls', any individuals with an 'eligible case' relative were excluded from the comorbidity sample. If another 'eligible control' was a relative, we randomly selected 1 member of the 'eligible control' pair as a 'control'. 'Eligible controls' without any relatives among 'eligible cases' or 'eligible controls' were selected as 'controls'. Age was defined as participant's age at recruitment into the UK Biobank study (Field #21022). Participant's biological sex was defined as sex determined from genotyping analysis (Field #22001).

Genotyping, quality control, and imputation was performed centrally by UK Biobank.[25] Among our UK Biobank study population, we filtered our genotype data to only include variants with MAF >1%, INFO score >0.8, and Hardy-Weinberg equilibrium exact test p-value great than 1e-6. We required a maximum per-variant and per-sample missing call rate < 0.1. Our final target dataset included 9.9M SNPs after excluding all duplicate SNPs.

3.3.2 Base Data: COVID-19 Host Genetics Initiative

We used data from the COVID-19 Host Genetics Initiative (COVID-19-HG), formed in response to the COVID-19 epidemic to generate, share, and analyze data to learn the genetic determinants of COVID-19 susceptibility, severity, and outcomes. Our base data was formed from the COVID-19-HG Round 5 meta-analysis released on January 18th, 2021. The 'COVID-19 Severity' base GWAS used the 'very severe respiratory confirmed COVID-19 vs. population controls' meta-analysis, including 4,792 cases and 1,054,664 controls of European ancestry, excluding UK Biobank participants. The specific case/control definition and counts are provided in **Table 5** below. The COVID-19 Severity GWAS included 9,944,485 SNPs.

3.3.3 Data Analysis

We used PRSice-2 to construct PRS for our 'severe COVID-19 infection' exposure of interest.[70] PRSs were constructed across a specific range of p-value thresholds: 5e-8, 1e-6, 1e-5, 1e-4, 1e-3, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, and 0.5 for the 28 Elixhauser comorbidity outcomes that had more than 1,000 cases. SNPs were clumped to obtain variants in linkage equilibrium with an $r^2 > 0.1$, p-value=1, or within 250kb to both ends of the index SNP. Previous PRS methods use computationally intensive permutations to evaluate the significance of a given PRS; and, select the optimal PRS p-value threshold. We used the less computationally intensive PRS principal component analysis (PRS-PCA) approach where we performed principal components analysis on the resulting set of PRSs and used the first PRS principal component (PRS-PC) in association tests with the 28 phenotypes of interest.[71] The first PC re-weights the variants included in the PRS to achieve maximum variation over all p-value thresholds used. Compared to the permutation optimization method, the PRS-PC method reduces type 1 error and overfitting. We then evaluated the association between the 'Severe COVID-19 Infection' PRS and each of the 28 Elixhauser comorbidities using logistic regression, adjusting for participant age, genetic sex, and the first 10 population structure principal components. Odds Ratios (ORs) and 95% CIs were calculated based on a 1 standard deviation (SD) increase in the PRS.

Variant filtering in both the base and target data was completed using Plink 2. PRSs were calculated using PRSice-2. Identification of the study population, Elixhauser comorbidities, covariates, descriptive analysis, PRS-PCA analysis, and logistic regression were completed using R version 3.6.1.[42]

3.4 Results

We constructed 31 case/control samples for each of the Elixhauser comorbidities from 502,493 participants in the UK Biobank cohort according to the methodology described in our methods section and displayed in **Figure 4**. Details on the total sample, # of cases, mean age, and sex distribution for each of the 31 Elixhauser comorbidity case/control samples is provided in **Table 6**.

The regression results for our 28 'severe COVID-19 infection' PRSs and the Elixhauser comorbidity outcome are provided in **Table 7** and **Figure 6**. We did not calculate PRSs for 3 Elixhauser

comorbidity samples given that they had very small numbers of cases ('*Anemia, blood loss*', '*Drug abuse*', and '*HIV/AIDS*'). We identified 4 significant associations between our 'severe COVID-19 infection' PRSs and our comorbidity phenotypes: '*Diabetes, uncomplicated*', '*Hypertension, uncomplicated*', '*Obesity'*, and '*Renal Failure*'. Among these 4 phenotypes, '*Hypertension, uncomplicated*' had the largest # of cases (n=107,604), followed by: '*Obesity*' (n=39,499), '*Diabetes, uncomplicated*' (n=34,523), and '*Renal failure*' (n=14,563). Our most significant association was observed for '*Obesity*' (p-value=1.93e-06), followed by '*Diabetes, uncomplicated*' (p-value=4.91e-05), '*Hypertension, uncomplicated*' (p-value=0.001), and '*Renal failure*' (p-value=0.015). Our strongest effect size (per 1 SD increase in PRS) was observed for our 'severe COVID-19 infection' PRS and '*Obesity*' (OR=1.03, 95% CI: 1.02-1.04), followed by: '*Diabetes, uncomplicated*' (OR=1.02, 95% CI: 1.01-1.04), '*Renal failure*' (OR=1.02, 95% CI: 1.00-1.04); and, '*Hypertension, uncomplicated*' (OR=1.02, 95% CI: 1.01-1.03).

3.5 Discussion

Using the largest GWAS results to date for severe COVID-19 infection, we identified shared genetic etiology between severe COVID-19 infection and 4 comorbidities ('*Diabetes, uncomplicated*', '*Hypertension, uncomplicated*', '*Obesity'*, and '*Renal Failure*') in the UK Biobank cohort. To date, this research represents the most comprehensive assessment of shared genetic risk for severe COVID-19 and EHR-derived comorbidities.

Our findings build upon recent research by the COVID-19 HGI on the genetic correlation and causal relationships between COVID-19 and other traits. Ganna et.al. used LD score regression to estimate genetic correlations between the COVID-19 HGI meta-analyses and GWAS summary statistics for 38 disease, health and neuropsychiatric phenotypes, selected based on their putative relevance to disease susceptibility, severity, or mortality.[72] They identified significant positive genetic correlations between COVID-19 critical illness and body mass index (BMI), Attention Deficit Hyperactivity Disorder, Coronary artery disease, Diabetes, Ischemic stroke, and Lupus. Three of these six significant findings have approximate synonyms with our Elixhauser comorbidity phenotypes and similar findings regarding

shared genetic etiology: 'BMI' and 'Obesity'; 'Coronary artery disease' and 'Hypertension, uncomplicated'; and, 'Diabetes' and 'Diabetes, uncomplicated' in our analysis. Our findings with respect to 'Renal failure' are not replicated in this analysis, although this could be due to key differences between our 'Renal failure' phenotype and their 'Chronic Kidney Disease' GWAS: 'Chronic Kidney Disease' would represent a less severe, broader phenotype than 'renal failure'. Application of two-sample MR to these 38 traits, excluding UK Biobank samples from their exposure meta-analysis, identified a significant causal association between genetically predicted higher BMI and COVID-19 critical illness (OR=1.4, 95% CI: 1.2-1.6).

There are a number of important differences between our research and the work by Ganna et.al. despite leveraging the same GWAS summary statistics for severe COVID-19 infection. First, selection of target phenotypes by Ganna et.al. was completed based on 'putative relevance' to COVID-19. Our target phenotypes were derived based on a standardized set of comorbidities, not researcher-informed selection; providing a more expansive list of target phenotypes across the disease phenome. Second, both LD-score regression and MR methods employed in their research leveraged pre-calculated summary statistics from a variety of cohorts and phenotypes. Consequently, the populations, methods, and data sources used to define 'cases' for a given disease will be heterogenous across the 28 sets of summary statistics used (e.g. the 'Heart Failure' meta-analysis[73] included cases identified via varying combinations self-report, physician diagnoses, biomarkers, imaging, and other sources across 28 contributing studies while the 'Chronic Kidney Disease' meta-analyses[74] was conducted using solely biomarker-based case definitions). In contrast, associations between our PRSs and target phenotypes were all based upon consistent case/control definitions, leveraging the same data sources, within the same cohort, across the disease phenome. In addition, while both LD-score-regression (used by Ganna et.al.) and PRS (used in our research) both can be exploited to identify shared genetic etiology among complex traits, our PRS approach provides an estimate of genetic liability to severe COVID-19 infection at the individual level.[24]

A number of other studies have identified obesity as a significant risk factor for both COVID-19 susceptibility and severity.[75–80] With respect to shared genetic relationships between obesity and COVID-

19, our results are consistent with findings from a two other recent UK Biobank studies. Using two-sample MR, *Aung et.al.* found that individuals in the highest genetic risk score quintiles of BMI (body mass index) were more susceptible to COVID-19.[81] *Zhu et.al.* identified a significant positive relationship between an individual's overall genetic risk for BMI (defined via PRS) and the risk of severe COVID-19 (defined as hospitalization following COVID-19 diagnosis).[80]

Outside of the work by Ganna et.al., we did not identify any published research specifically examining shared polygenic relationships between COVID-19 and hypertension, diabetes, and renal failure. Other studies can provide some additional biological context for our findings. Kasela et.al.'s analysis of genetic and non-genetic factors affecting the expression of COVID-19-relevant genes in the large airway epithelium found that obesity, hypertension, cardiovascular disease, and age were associated with COVID-19-relevant immunosuppression at the airway epithelium, which may contribute to both increased COVID-19 susceptibility and disease severity.[82] A study examining 'phenome-wide' EHR diagnoses in Michigan Medicine identified that severe COVID-19 (hospitalization + intensive care unit or death) had strong associations with circulatory system, genitourinary (renal diseases in particular), and respiratory diseases.[83] A similar study conducted in another integrated health system (Geisinger) examining the association between 21 clinical phenotypes and COVID-19 hospitalization also identified increased COVID-19 hospitalization risk for hypertension, diabetes, and renal failure[84] This study highlighted both Stage IV chronic kidney disease (CKD) and end-stage renal disease (Stage V CKD) as the strongest associations with COVID-19 severity.

There are a few important limitations to our research. First, we did not investigate causal relationships between polygenic risk for severe COVID-19 infection and our Elixhauser comorbidity phenotypes. The largest difference between our research and the causal investigations via two-sample MR by Ganna et.al. is with respect to our selection of disease phenotypes; and, our consistent application of case/control definitions across phenotypes. There are no published GWAS summary statistics available that use the Elixhauser comorbidity definitions. Consequently, future work stemming from these initial efforts will be required to conduct at least 28 novel GWAS utilizing these case/control

definitions to generate the summary statistics required for further causal investigation using MR. Second, our analysis was limited only to only individuals of European ancestry. Unfortunately, our current data sources do not have sufficient racial representation to investigate our primary hypothesis. Consequently, the genetic inferences identified in our data may not be valid for populations of non-European ancestry. Third, cases identified via primary care data were ascertained only from TPP EHR data. The absence of a published, validated code list for the Elixhauser Comorbidities in SNOMED-CT prevented our inclusion of cases identified in EMIS practices in the past 20 years. Given the similarities in patient populations between TPP and EMIS practices, we do not believe the exclusion of cases from EMIS practices would have resulted in biased results. Rather, the additional cases we could have gained from identification of cases and controls in EMIS practices may have resulted in significantly larger sample sizes. Fourth, our definition of Elixhauser Comorbidities was derived entirely from diagnosis codes. Finally, there is significant potential for within-disease heterogeneity due to coding practices by different health care providers; as well as vulnerabilities due to provider ascertainment of disease warranting recording a diagnosis code. It is unclear how these coding practices may have impacted our genetic findings as the impact of coding practices (and generally case ascertainment) has not been well studied in the context of polygenic risk scores.

There are notable strengths in our research. COVID-19 studies that leverage UK Biobank data frequently use self-reported conditions, biomarkers, and other variables that were measured during the Baseline Assessment visit between 2006 and 2010. Each of our conditions was measured via diagnosis codes, which required a healthcare provider to record the diagnosis based on a clinical event and evaluation. Furthermore, we did not need to translate any of our diagnosis code lists across medical vocabularies (e.g. from CTV-3 to SNOMED-CT; or ICD-10 to CTV3), potentially losing or modifying the original clinical meaning of the diagnosis. Our code lists for the Elixhauser comorbidities were developed specifically for each medical vocabulary. We also did not select specific comorbidities for investigation based on our own opinions, information from prior research, or with the intention to validate a specific hypothesis. Rather, the use of a common measure of comorbidities ensured that the

comorbidities we included for investigation were comprehensive. Finally, our study was able to leverage the UK Biobank's primary care data linkage for COVID-19 investigation. This ensured complete ascertainment of 'cases' within the UK healthcare system; and, confidence that our control group was disease free.

Our research indicates that the same genetic composition that increases an individual's risk for COVID-19 may also influence their risk for other important comorbid diseases. Future research, both genetic and purely observational studies, will be required to validate the relationships that we have described in our research. We hope that the associations we have described here will provide a solid foundation for these investigations, especially as new data emerges, and additional longitudinal data accumulates.

All UK Biobank data was accessed in accordance with GlaxoSmithKline's UK Biobank Application #20361.

Chapter 3, in full, is currently being prepared for submission for publication of the material. Davitte, J.; Salem, R.; Pines, H.; Martin, N.; Brodine, S.; Shaffer, R. The dissertation author was the primary investigator and author of this material.

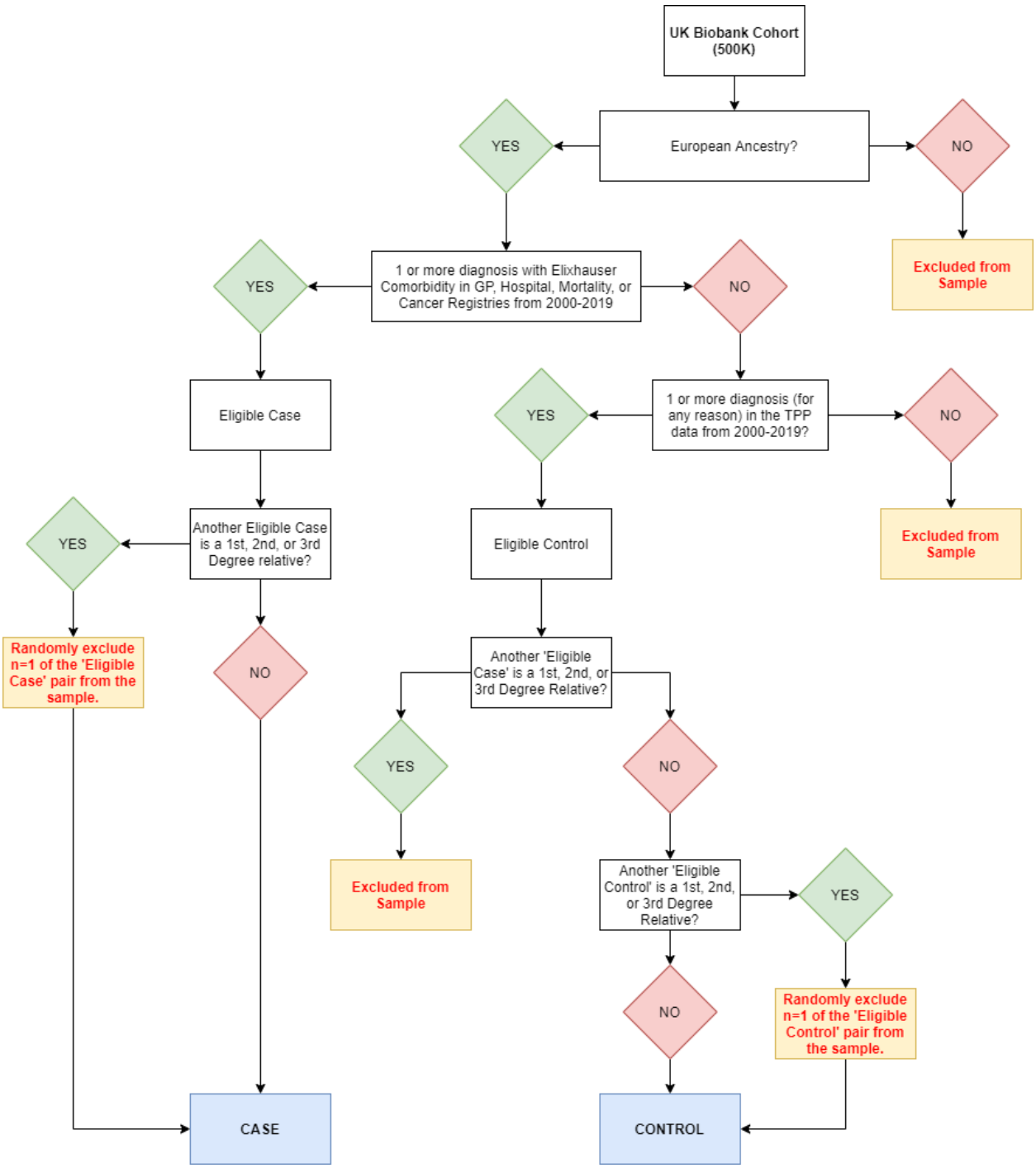


Figure 4: Case/Control Selection Algorithm for Each of the 31 Elixhauser Comorbidity samples.

Table 5: Case/Control Definitions for Base Severe COVID-19 GWAS from COVID-19 Host Genetics Initiative Round 5 Meta-Analysis

	Severe COVID-19 Infection
COVID-19-HG Analysis	A2_ALL_eur_leave_ukbb
Case Definition	Required to meet **ALL** of the following: 1) Laboratory confirmed SARS-CoV-2 infection 2) Hospitalized for COVID-19 3) Death or Respiratory Support
Control Definition	Everyone that is not a case, e.g. population
Total Cases	5,870
Total Controls	1,054,664

Table 6: Descriptive Frequencies for Case/Control Definitions by Elixhauser Comorbidity Phenotype, UK Biobank Cohort

Elixhauser Comorbidity	**Total Sample**	**# of Cases**	**Age** *Mean (SD)*	**Male** *N (%)*
Alcohol abuse	137,296	15,106	57.0 (7.9)	65,103 (47.4)
Anemia, deficiency	137,319	16,243	57.1 (7.9)	63,091 (45.9)
Anemia, blood loss*	128,570	602	57.0 (7.9)	59,496 (46.3)
Cardiac arrhythmia	152,111	41,601	57.7 (7.9)	74,283 (48.8)
Congestive heart failure	135,729	12,809	57.3 (7.9)	64,348 (47.4)
Coagulopathy	130,854	4,564	57.0 (7.9)	60,794 (46.5)
Depression	141,370	26,780	56.9 (7.9)	63,998 (45.3)
Diabetes, complicated	130,785	6,322	57.1 (7.9)	61,193 (46.8)
Diabetes, uncomplicated	146,112	34,523	57.4 (7.9)	70,368 (48.2)
Drug abuse*	128,762	800	57.0 (7.9)	59,740 (46.4)
Fluid/electrolyte disorders	139,069	18,683	57.4 (7.9)	64,987 (46.7)
HIV/AIDS*	128,153	175	57.0 (7.9)	59,367 (46.3)
Hypertension, complicated	129,676	2,406	57.0 (7.9)	60,332 (46.5)
Hypertension, uncomplicated	191,233	107,604	58.3 (7.7)	93,677 (49.0)
Hypothyroidism	141,454	22,358	57.2 (7.9)	61,581 (43.5)
Liver disease	134,079	10,094	57.1 (7.9)	62,457 (46.6)
Lymphoma	130,772	4,281	57.0 (7.9)	60,759 (46.5)
Metastatic cancer	138,799	16,822	57.3 (7.9)	64,238 (46.3)
Other neurological disorders	137,168	18,217	57.2 (7.9)	64,161 (46.8)
Obesity	143,080	39,499	57.1 (7.9)	66,245 (46.3)
Paralysis	130,436	3,762	57.0 (7.9)	60,615 (46.5)
Peptic ulcer disease	133,230	8,914	57.1 (7.9)	62,090 (46.6)
Peripheral vascular disease	135,429	12,330	57.3 (7.9)	64,441 (47.6)
Psychoses	129,227	1,919	57.0 (7.9)	59,833 (46.3)
Chronic pulmonary disease	157,385	56,177	57.3 (7.9)	72,458 (46.0)
Pulmonary circulation disorder	133,154	7,974	57.1 (7.9)	62,099 (46.6)
Rheumatoid arthritis	136,044	14,171	57.2 (7.9)	61,819 (45.4)
Renal failure	136,963	14,563	57.4 (7.9)	63,920 (46.7)
Solid tumor without metastasis	161,395	54,796	57.7 (7.8)	75,657 (46.9)
Valvular disease	136,984	14,191	57.3 (7.9)	64,502 (47.1)
Weight loss	134,105	10,448	57.1 (7.9)	61,990 (46.2)

***** *These 3 conditions were excluded from the PRS analyses due to less than 1,000 cases identified.*

Table 7: Regression Results for Severe COVID-19 Infection Polygenic Risk Scores (PRS) Against 28 Elixhauser Comorbidity Outcomes

Elixhauser Comorbidity Outcome	ß (SE)[a]	OR (95% CI)[b]	P-value
Alcohol abuse	-0.011 (0.009)	0.989 (0.972,1.006)	0.19
Anemia, deficiency	-0.007 (0.008)	0.993 (0.977,1.010)	0.422
Cardiac arrhythmia	-0.005 (0.006)	0.995 (0.983,1.007)	0.402
Congestive heart failure	0.004 (0.010)	1.004 (0.985,1.023)	0.675
Coagulopathy	-0.023 (0.015)	0.977 (0.948,1.006)	0.124
Depression	0.010 (0.007)	1.010 (0.996,1.024)	0.147
Diabetes, complicated	0.014 (0.013)	1.014 (0.988,1.040)	0.287
Diabetes, uncomplicated	**0.026 (0.006)**	**1.026 (1.013,1.039)**	**4.91e-05*** **
Fluid/electrolyte disorders	0.001 (0.008)	1.001 (0.986,1.017)	0.864
Hypertension, complicated	0.016 (0.021)	1.016 (0.975,1.058)	0.451
Hypertension, uncomplicated	**0.016 (0.005)**	**1.016 (1.006,1.026)**	**0.001****
Hypothyroidism	0.003 (0.008)	1.003 (0.988,1.018)	0.739
Liver disease	0.010 (0.010)	1.010 (0.989,1.031)	0.356
Lymphoma	-0.006 (0.016)	0.994 (0.964,1.025)	0.718
Metastatic cancer	0.000 (0.008)	1.000 (0.983,1.016)	0.978
Other neurological disorders	-0.007 (0.008)	0.993 (0.977,1.009)	0.392
Obesity	**0.028 (0.006)**	**1.029 (1.017,1.041)**	**1.93e-06*** **
Paralysis	-0.002 (0.017)	0.998 (0.966,1.031)	0.922
Peptic ulcer disease	0.013 (0.011)	1.013 (0.991,1.035)	0.259
Peripheral vascular disease	0.016 (0.010)	1.016 (0.996,1.035)	0.11
Psychoses	-0.019 (0.023)	0.981 (0.937,1.026)	0.401
Chronic pulmonary disease	0.004 (0.005)	1.004 (0.993,1.015)	0.467
Pulmonary circulation disorder	0.007 (0.012)	1.007 (0.985,1.031)	0.53
Rheumatoid arthritis	0.008 (0.009)	1.008 (0.991,1.026)	0.361
Renal failure	**0.022 (0.009)**	**1.022 (1.004,1.041)**	**0.015***
Solid tumor without metastasis	0.004 (0.005)	1.004 (0.993,1.015)	0.479
Valvular disease	0.007 (0.009)	1.007 (0.989,1.025)	0.475
Weight loss	0.000 (0.010)	1.000 (0.980,1.021)	0.994

[a] *Regression coefficient for 1st principal component of severe COVID-19 PRS against each Elixhauser Comorbidity outcome; adjusting for age, genetic sex, and the first 10 principal components of population structure.*

[b] *Odds per 1 SD increase in PRS*

** p-value < 0.05; ** p-value < 0.01; *** p-value < 0.001*

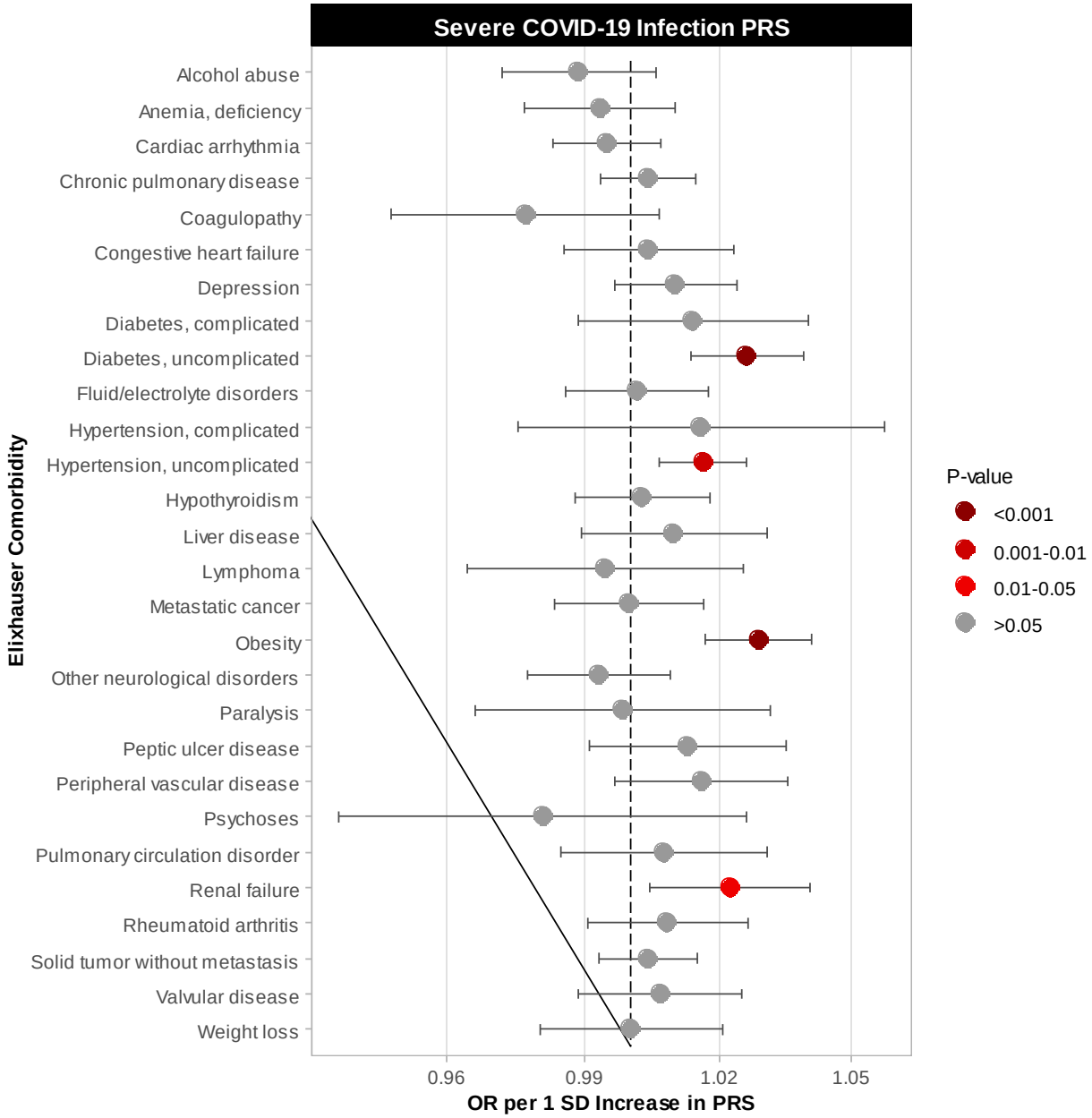


Figure 5: Odds Ratios and 95% Confidence Intervals for Severe COVID-19 Polygenic Risk Scores (PRS) Against 28 Comorbidity Outcomes, UK Biobank

Chapter 4: Proximity to Local Military Bases and HIV Infection Among Adolescent Girls and Young Women Living in Communities Surrounding Military Bases in Mozambique

4.1 Abstract

In Mozambique, local military populations may be important members of local sexual networks. From July 2018 to January 2019, 8,034 women between the ages of 15 and 35 were recruited for a cross-sectional survey from venues surrounding four military bases in Mozambique where adolescent girls and young women (AGYW) congregate or meet sexual partners. Study personnel captured the Global Positioning System (GPS) location where the female participant was recruited, administered a behavioral survey collecting demographics and HIV risk behaviors, and provided HIV testing to assess current HIV status. Geospatial methods were used to estimate travel time to the nearest military base (accounting for travel speed, roads, and surface types) from the locations where female participants were recruited. We used multivariable logistic regression to calculate adjusted odds ratios (ORs) and 95% confidence intervals (95% CIs) for HIV-positive status for each 15 minute increase in our travel time exposure, adjusting for participant age, current marital status, education, potential hazardous drinking, and transactional sex. Our final sample was comprised of 7,514 AGYW with an overall HIV prevalence of 6.3% (n=473/7,514). Among participants across all 4 bases combined, the odds of HIV-positive decreased for each 15 minute increase in travel time (OR=0.91, 95% CI: 0.86-0.96) after adjusting for covariates. While the effect was in the same direction, there was no significant difference in the odds of incident HIV-positive diagnosis for increasing travel time (OR=0.95, 95% CI: 0.89-1.01). For the results stratified by Base, there were no significant findings for Base 1, Base 2, or Base 3 for either of our outcomes. However, the odds of both HIV-infection (OR=0.77, 95% CI: 0.62-0.96) and incident HIV infection (OR=0.73, 95% CI: 0.57-0.93) were significantly lower for each 15 minute increase in travel time among participants in Base 3. While our findings support the hypothesis that AGYW congregating or meeting sexual partners at venues in closer proximity to military bases is positively associated with HIV infection in our overall sample, our stratified analysis indicate that this hypothesis does not hold true across all types of military bases and the communities that surround them.

4.2 Introduction

Eastern and Southern Africa remains the region most affected by the human immunodeficiency virus (HIV) epidemic: accounting for 45% of the world's new HIV infections and 53% of people living with HIV (PLHIV) globally.[85] In 2017, an estimated 800,000 [650,000-1,000,000] people in Eastern and Southern Africa acquired HIV and 380,000 [300,000 – 510,000] people died of AIDS-related illness, with Mozambique, South Africa, and the Republic of Tanzania accounting for more than half of new HIV infections and deaths from AIDS-related illness in the region.

According to 2017 estimates from UNAIDS, Mozambique accounts for 16% of all new HIV infections and 18% of all AIDS-related deaths in eastern and Southern Africa.[85] In Mozambique, UNAIDS estimates an overall HIV prevalence of 12.3% and an HIV incidence of 6.60 per 1,000 among adults ages 15 to 49 years.[86] The Demographics and Health Survey (DHS) in Mozambique, Inquérito de Indicadores de Imunização, Malária e HIV/SIDA em Moçambique (IMASIDA) in 2015, revealed that HIV prevalence is generally higher among women than men (15.4% vs. 10.1%) and increasing with age, reaching a peak between the ages of 35-39 for both men (18%) and women (25%).[87]

Collectively, the Army, Air Force, and Maritime Wing Branches constitute the *Forças Armadas de Defesa de Moçambique* (FADM). In 2016, the FADM conducted their third HIV/AIDS Seroprevalence and Behavioral Epidemiology Risk Survey (SABERS) in collaboration with the United States Department of Defense HIV/AIDS Prevention Program (DHAPP), United States Embassy in Mozambique, and Research Triangle International (RTI).[88] HIV prevalence data in foreign militaries is frequently kept private as many countries continue to perceive military HIV data as a national security issue. However, in July 2018, the National Director of Military Health in Mozambique for the Forças Armadas de Defesa de Moçambique (FADM) publicly released the military HIV prevalence of 12.6% in 2016, based on the results from their HIV/AIDS Seroprevalence and Behavioral Epidemiology Risk Survey (SABERS).[88,89]

While the overall HIV prevalence of 12.6% among active-duty military personnel was slightly lower than the overall 13.2% HIV prevalence in the general adult population (as measured in the 2015

IMASIDA survey), HIV prevalence among military men was significantly higher than adult men in the general population: 13.2% compared to 10.1%.[87,88] While specific data on sexual risk behaviors in the FADM are not publicly available; circumstances within the military environment, including high mobility, long periods away from home, disposable income, and greater numbers of casual sexual relationships amplify the risk for both HIV contraction and transmission.[90–93] Extended, foreign deployments are common, where military personnel live and interact freely with local populations. Soldiers are often younger, susceptible to peer pressure, sexually active, and report high rates of risky sexual behavior and low condom use.[94–99] Multiple HIV phylogenetic studies in SSA have shown that communities "export' and "import" strains, demonstrating that mobility-driven transmission frequently occurs and may be important contributors to local HIV epidemics.[100–103] Consequently, given the high prevalence of HIV among FADM men compared to the general population of men, known HIV risk behaviors common to military populations, and sustained mobility across Mozambique, FADM military men and the bases where they are stationed may represent important components of local sexual networks.

Adolescent girls and young women (AGYW), aged 15-24, are one of the most important sub-populations for HIV prevention efforts. HIV disproportionately affects adolescent girls and young women due to unequal cultural, social, and economic status in their communities.[85] In sub-Saharan Africa, adolescent girls and young women accounted for one in four HIV infections in 2017 despite comprising only 10% of the population.[85] In Eastern and Southern Africa, young women acquire HIV five to seven years earlier than their male peers, often synonymously with sexual debut.[104] At the population level, the high incidence of HIV in adolescent girls and young women is a key component sustaining intergenerational transmission of HIV.[105]

Socio-behavioral factors are important drivers of HIV vulnerability in adolescent girls and young women.[104] A study characterizing AGYW and their sexual partners in multiple districts of Mozambique found that relationships between AGYW and their male partners were characterized by high risk for HIV; reporting multiple sexual relationships, regardless of marital status.[106] In addition, most men reported

using condoms inconsistently. Condom use was more commonly reported in sexual partnerships with AGYW who were younger (15-19 years) rather than older (20-24 years). AGYW were less willing to negotiate condom use in relationships where the male partner provided money or other benefits; highlighting how gender norms and power dynamics can create barriers to condom negotiation. Numerous other studies have shown that power dynamics within sexual relationships contribute to high rates of HIV and other sexually transmitted infections (STIs).[107–111] Furthermore, AGYW are particularly vulnerable to coercive economic circumstances, leading to 'transactional sex', defined as non-marital, non-commercial sexual encounters or relationships primarily motivated by the implicit assumption that sex will be exchanged for material benefit or status.[112] Given that 45% of the total population in Mozambique is under the age of 15, the population of adolescent girls and young women at significant risk for HIV infection will continue to grow as girls reach age 15 and beyond.[113]

It is unclear how local military populations, stationed at bases located in a mix of urban and rural settings, engage in local sexual networks, particularly in sexual networks comprised of primarily AGYW. Unfortunately, direct research into sexual relationships between active-duty military men and women living in communities surrounding bases is infeasible. Women may not be aware whether sexual partner is a member of the military, especially for casual partners or sex work clients. These men may be in plain clothes when not present within the confines of the military base; and, they may be especially careful not to wear military attire when frequenting bars, nightclubs, markets, or sex work venues. In addition, particularly for younger women and/or women engaging in transactional sex, these women may be hesitant to report military sexual partners due to the gender-power dynamics discussed above.

This study aimed to examine the association between proximity to the local military base (measured via travel time to the nearest base from AGYW recruitment location) and HIV infection among AGYW (15-35 years of age) living in communities surrounding military bases in Mozambique. Given the challenges with directly assessing the presence of sexual relationships between military men and young women, we have selected proximity to military facilities as an indicator of accessibility to military sexual partners. These military bases have high concentrations of men; and, are frequently surrounded by

venues (e.g. bars, nightclubs, markets, etc.) where military personnel meet sexual partners. We hypothesized that congregating or meeting sexual partners at venues in closer proximity to military bases is positively associated with HIV infection.

4.3 Methods

4.3.1 Study Setting and Population

From July 2018 to January 2019, 8,034 women between the ages of 15 and 35 were recruited for a cross-sectional survey from communities surrounding four military bases selected as study sites for the Treat All Study in Mozambique, a PEPFAR-funded study to evaluate; 1) the effects of rapid scale-up of ART in military bases with high HIV prevalence; and, 2) HIV transmission dynamics between the military and the communities surrounding military bases.

Four military bases and their surrounding communities were included as part of this study. Due to national security sensitivities, the precise locations of the bases used in this analysis will not be disclosed. For this reason, we have de-identified the military bases and labeled each base as "Base 1", "Base 2", "Base 3", and "Base 4". Each of the selected military bases are located in different provinces of Mozambique: one in the north, two in central, and one in western Mozambique. Base 1 is located within 5km of the most densely populated urban center of all the selected bases. Base 2 is located in a rural area with limited access to the closest urban center via a difficult to navigate, primarily dirt road. Base 3 and Base 4 are also located just outside of urban centers; however, with better road access to the central city compared to Base 2 and less densely populated urban centers compared to Base 1. Bases 3 and 4 have the largest numbers of troops stationed on their premises. Base 1 had slightly less troops stationed compared to Bases 3 and 4; however, it had significantly more troops compared to Base 2, which had by far the smallest number of troops.

Before initiating each cross-sectional survey, study personnel created community maps to guide recruitment. Each of the four community maps was developed in collaboration with the district Ministry of Health, FADM, and the Conselho Nacional de Combate ao HIV/SIDA Moçambique (CNCS). Each community map to identified important locations where adolescent girls and young women (ages 15-24)

and women (ages 25-35) congregate or meet sexual partners, including venues such as markets, bars, nightclubs, schools, and bus stops. Route plans were generated for each research team based on the estimated target AGYW population size for a given portion of the community map; and, were adjusted to include day and/or night recruitment, depending on the findings at specific hot spots. From July 2018 to January 2019, nurse research officers followed the previously developed route plans and began study recruitment. Once at a targeted location, the nurse research officers would introduce him/herself to potential AGYW participants, including their name, role, and credentials. The nurse research officer would then invite the participant to a private area where they could assess study eligibility and complete study procedures. AGYW participants were eligible if they were 15 to 35 years of age, consented to completed an HIV survey and HIV testing; and, if they tested HIV-positive, would consent to a blood draw for HIV genetic sequencing and assessment of recent infection. Following eligibility assessment and informed consent, the research nurse recorded the GPS location where the participant had been recruited and administered a survey using tablet devices, collecting demographics and HIV risk behavior information from the participant.

All cross-sectional survey data, geolocation data, and HIV-testing results were captured within the Research Electronic Data Capture (REDCap) system, a secure web application for building and managing electronic surveys and databases, specifically designed to support online and offline data capture for research studies and operations.[114]

4.3.2 Exposure of Interest: Travel Time to Nearest Military Base

Locations for each of the four military bases were recorded by study staff using global positioning system (GPS) coordinates. GPS coordinates were also captured by each study staff member at the location where each female participant was recruited.

All geospatial analysis was completed using ArcGIS Pro version 2.7 software.[115] A basemap layer for the lakes, seas, oceans, large rivers, and dry salt flats was retrieved from Esri.[116] Country and provincial boundaries were retrieved from the "GADM database of Global Administrative Areas".[117] Polyline data for roads and road types in Mozambique were retrieved from the OpenStreetMap (OSM)

database.[118] For each road segment, "highway" and "surface" attributes were used to categorize the segment into one of five distinct road types: paved major road, unpaved major road, paved minor road, unpaved minor road, and trail. All geospatial data was projected into Moznet UTM zone 37S coordinates to ensure that all measurements were in units of meters.

Participants' proximity to the nearest military base was defined as the least accumulated travel time (henceforth referred to as 'Travel Time') between the GPS coordinates for their recruitment location and the closest military base and calculated using the ArcGIS cost distance tool.[116] Euclidean distance measures use the shortest distance between Point A and Point B (a straight line). These distance measures do not account for landscape, terrain, roads, or other travel surfaces that may be important factors that impact how a person might travel from Point A to Point B. Cost distance tools are similar to Euclidean tools, but instead of calculating the linear distance from one location to another, they determine the least accumulated travel cost from each point to the nearest source location, applying distance in cost units, not in geographic units.

We assumed that women would walk to the nearest major road to access public transportation, which would increase their travel speed on these surfaces. Speeds for each road and surface type were defined using the estimates for travel in dry weather according to the methodology described by Makanga, et.al..[119] We used these speed estimates to generate a raster from the OSM roads dataset using the "Convert Polyline to Raster" tool to capture the time in minutes required to travel 5 meters for each $5m^2$ road segment. We assumed that travel by water was uncommon, generating a separate raster for bodies of water that assigned a value of "99,999" for the time in minutes required to travel for each $5m^2$. We generated a cost surface by combining the roads and water bodies rasters, assigning any non-road and non-water surfaces a value of 0.1 minutes per $5m^2$ cell (equivalent to walking on general terrain). We have provided the estimates used to generate the cost surface in **Table 8** for reference. The "Cost Distance" tool was used to generate a cost distance raster using the military base coordinates as the source location and the cost surface raster that was generated in the previous step. We used the "Extract Values to Points" tool to assign a value for the accumulated travel time in minutes from the cost distance raster

for each participant's location. These values, representing the least accumulated cost of travel (measured in minutes) from the participant's recruitment location and the nearest military base was used as our 'Travel Time' exposure.

4.3.3 Outcome of Interest: HIV Infection

Following completion of the cross-sectional survey, participants were asked whether they had a documented HIV-positive status and, if HIV-positive, whether they were on antiretroviral therapy. 'Documents' for HIV-positive status encompassed a variety of documents, e.g. ART clinic cards, referrals from HIV testing sites, or other documents generated from previous HIV testing. For participants with unknown HIV status, HIV negative status, undocumented HIV-positive status, or a documented HIV positive status but not currently taking ART, the research nurse conducted HIV testing according to the national HIV testing algorithm. National guidelines at the time of the study required the use of a combination of Determine and Unigold HIV rapid test kits. A non-reactive Determine test led to a final HIV-negative status. A reactive Determine test followed by a reactive Unigold test led to a final HIV-positive status. For reactive Determine and non-reactive Unigold tests, the tests were repeated. If the tests did not agree again, then the client was indicated as having an "indeterminate" HIV status. Blood samples were collected from participants with documented HIV-positive status and on ART in addition to any participants that were newly tested as HIV-positive. For HIV-positive participants that were not on ART, counselors provided referral to a preferred location for ART evaluation and initiation. We additionally defined 'newly-diagnosed' HIV-positive participants as participants that tested HIV-positive in our study procedures, but who indicated in our cross-sectional survey (prior to HIV testing) they were HIV-Negative, didn't know their HIV status, or preferred not to answer our HIV testing question.

4.3.4 Covariates

Variables that were related to both HIV-status and congregating or meeting sexual partners in locations that are more geographically accessible to military bases were explored as potential covariates for this analysis. We used directed acyclic graphs (DAGs) to select covariates to be included in our regression models (**Figure 6**). Chosen covariates included: participant age in years, current marital status

(single, single and living with a partner, married or in union, or other (polygamous marriage, divorced, widowed), highest education achieved (some primary or none, primary, or secondary or higher), potential hazardous drinking (yes, no), transactional sex with any of the three most recent sexual partners (yes, no). Potential hazardous drinking was measured by an Alcohol Use Disorders Identification Test (AUDIT-C) score of 3 or greater.[120] Each woman was asked whether they had ever received money, shelter, food, drugs, favors, or gifts in exchange for sex with each of their three most recent sexual partners. Transactional sex "Yes" was assigned for women who responded "Yes" to this question for any of their three most recent sexual partners.

Statistical Analysis

All non-geospatial statistical analysis was completed using R version 3.6.1.[42] We calculated descriptive frequencies for our outcome variables and covariates, both overall, and stratified by military base. In addition, we calculated additional descriptive frequencies for these variables with additional stratification of each military base sample by the HIV-status of female participants. We used directed acyclic graphs to select covariates for our multivariable logistic regression models. We used multivariable logistic regression to calculate adjusted odds ratios (ORs) and 95% confidence intervals (95% CIs) for a 15 minute increase in our 'Travel Time' exposure, adjusting for participant age, current marital status, education, potential hazardous drinking, and transactional sex. Finally, we conducted an additional analysis by limiting our outcome to only participants who were newly diagnosed with HIV at the time of the cross-sectional survey. We used HIV-negative status as the reference category for both outcomes.

4.4 Results

Of the 8,034 women that were approached to participate in our study, 7,971 consented to participate. We additionally excluded 426 women due to inconsistencies in data collection, 12 with missing or indeterminate HIV testing results, and 19 with missing or impossible GPS locations. Our final sample for this study was comprised of 7,514 AGYW living in the communities surrounding the four selected military bases.

Table 9 provides participant characteristics both overall and by the nearest military base. The overall HIV prevalence in our sample was 6.3% (n=473/7,514). Of the n=473 HIV-positive participants, 67.9% (n=321/473) were newly diagnosed with HIV at the time of the study. Participants were: on average 23.4 (SD=5.5) years old; mostly single, never married, and not living with a partner (42.3%, n=3,175); or married or in a union (30.4%, n=2,283). The majority of participants had completed primary school (57.9%, n=4,163), followed by 22.5% (n=1,619) having completed secondary school or higher and 19.6% (n=1,414) having not completed primary school (or had no education). In addition, 11.5% of participants (n=862) screened positive for hazardous drinking according to the AUDIT-C. Regarding transactional sex, 7.6% (n=573) of our sample reported receiving money, shelter, food, drugs, favors, or gifts in exchange for sex with at least one of their three most recent sexual partners. Finally, very few AGYW reported having had a member of the military as any one of their three most recent sexual partners (3.9%, n=296).

Participants were well distributed across the four bases: 36.9% (n=2,772) in Base 1, 23.4% (n=1,759) in Base 2, 18.4% in Base 3 (n=1,386); and, 21.3% in Base 4 (n=1,597). Base 4 had the highest HIV prevalence in the study sample at 8.5% (n=136/1597) followed by: 6.3% (n=88/1,386) in Base 3, 5.8% in Base 1 (n=161/2,772), and 5.0% (n=88/1,759) in Base 2. Participant age was similar across the bases. The composition of marital status was similar between Base 1 and Base 2, with most participants reporting that they were single, never married, and not living with a partner: 48.0% (n=1,329) and 47.9% (n=843) respectively. Base 3 had the highest proportion of participants that indicated that they were married (42.4%, n=588). Base 4 had the highest proportion of participants that reported being single and living with a partner (34.9%, n=557). Base 4 AGYW were the most educated: 33.5% (n=489) had 'Secondary or Higher') education compared to 24.0% (n=641) in Base 1, 17.1% (n=235) in Base 3, and 15.0% (n=15.0%) in Base 1. AGYW proximal to Base 3 were the least educated compared to the other bases: 26.8% (n=369) with primary school or no education compared to 23.5% (n=399) at Base 2, 18.6% (n=497) at Base 1, and 10.2% (n=149) at Base 4. The proportion of participants that screened positive for hazardous drinking was highest in Base 3 (13.2%, n=183) followed by Base 1 (8.4%, n=233), Base 4

(5.8%, n=93), and Base 2 (3.6%, n=64). Finally, transactional sex was most commonly reported in Base 3 (13.2%, n=183) compared to 8.4% (n=233) for Base 1, 5.8% (n=93) for Base 4, and 3.6% (n=3.6%) for Base 2. The proportion of participants that reported a military sexual partner (among their last 3 partners) was less than 10% across all four bases.

Participants were on average less than an hour of travel time away from the closest military base (Mean=53.3 minutes, SD=28.5). However, there was substantial variation in travel time across the study sites. The shortest travel time was observed among participants surrounding Base 4 (mean=30.3 min, SD=17.6), followed by 45.5 minutes (SD=22.6) for Base 1 and 57.1 minutes (SD=15.6) for Base 3. Participants located near Base 4 were substantially further from the military base with an estimated travel time of 83.5 minutes (SD=26.2). A histogram for the distribution of our travel time exposure is provided for the 4 bases in **Figure 7**.

Tables 10 and 11 provide a summary of participant characteristics by their current HIV-status and the nearest military base. HIV-positive participants were older on average across all bases. With regards to marital status, the majority of HIV-positive participants indicated that they were married or in a union across all sites: 43.5% (n=70) for Base 1, 36.4% (n=24) for Base 2, 44.3% (n=39) for Base 3, and 25.0% (n=34) for Base 4. By comparison, most HIV-negative participants reported being single, never married, and not living with a partner across all bases: 50.1% (n=1,308) for Base 1, 49.9% (n=833) for Base 2, 42.9% (n=557) for Base 3, and 27.9% (n=408) for Base 4. Hazardous drinking was consistently higher among HIV-positive participants at all bases. While transactional sex among participants was more common among HIV-positive women at Base 1, Base 2, and Base 3, more HIV-negative women reported transactional sex (6.1%, n=89) compared to HIV-positive women (2.9%, n=4) at Base 4. Finally, across all bases, most HIV-positive women were newly diagnosed through HIV-testing provided as part of our study procedures: 71.4% (n=115) at Base 1, 72.7% (n=64) at Base 2, 80.7% (n=71) at Base 3, and 52.2% (n=71) at Base 4.

ORs and 95% CIs calculated in our multivariate logistic regression models for each 15 minute increase in travel time for both comparison groups (HIV-positive vs HIV-negative and incident HIV-

positive vs HIV-negative) are provided in **Figure 8**. Among all participants, the odds of HIV-positive decreased for each 15 minute increase in travel time (OR=0.91, 95% CI: 0.86-0.96) after adjusting for covariates. While the effect was in the same direction, there was no significant difference in the odds of incident HIV-positive diagnosis for increasing travel time (OR=0.95, 95% CI: 0.89-1.01). For the results stratified by Base, there were no significant findings for Base 1, Base 2, or Base 3 for either of our outcomes. However, the odds of both HIV-infection (OR=0.77, 95% CI: 0.62-0.96) and incident HIV infection (OR=0.73, 95% CI: 0.57-0.93) were significantly lower for each 15 minute increase in travel time among participants in Base 3.

4.5 Discussion

Our study found that AGYW that meet or congregate near military bases were at a slightly elevated risk for HIV-infection in the combined sample. In our stratified analysis by military base, we observed a strong relationship between HIV-positive diagnosis (both overall and in the subset of newly diagnosed AGYW) in only 1 of our 4 military bases (Base 3). While our findings support the hypothesis that AGYW congregating or meeting sexual partners at venues in closer proximity to military bases is positively associated with HIV infection in our overall sample, our stratified analysis indicate that this hypothesis does not hold true across all types of military bases and the communities that surround them.

Base 1 is co-located with one of the most densely populated urban centers in Mozambique; with substantial access and availability of public transportation throughout the community. The numbers of military personnel stationed on Base 1 comprise an extremely small proportion of the overall community of men living within the urban center. This likely dilutes the likelihood that local women are engaging in sexual intercourse with an HIV-positive military man compared to a civilian HIV-positive man. This is supported by our finding that a substantially smaller proportion of AGYW in Base 1 reported having had a military sexual partner (2.0%) compared to more than 4% of AGYW in each of the other three bases. In addition, the availability of good road and public transportation infrastructure disperses the attendance of military men across a larger amount of venues where they may meet sexual partners (e.g. bars, nightclubs, markets, etc.). In this scenario, our hypothesis that proximity to military facilities is an indicator of

accessibility to military sexual partners may not be true given that the military population may be dispersed across a much wider geographic area and comprise a much smaller proportion of the overall HIV-positive male population. These features may offer some insight into the null association that we measured in our study for Base 1.

Base 4 is also located near the center of an urban area with substantial road access, although with a much smaller surrounding civilian population than Base 1. This location accounts for the low travel time between where AGYW were recruited and the Base 4's location (less than 30 minutes on average). This proximity to Base 4 may account for the higher proportion of AGYW that reported a military sexual partner (6.4%) compared to all other bases. However, AGYW at Base 4 were substantially more educated than Base 1 (33.5% 'Secondary or Higher' vs. 24.0%) and were less likely to engage in hazardous drinking (3.4% vs. 17.2%) or transactional sex (5.8% vs. 8.4%). These factors reflect lower sexual risk behaviors among AGYW at Base 4 compared to Base 1, which may have offset any risk related to engaging with military sexual partners.

Base 2, by comparison to the other three bases included in our study, has the smallest military force size stationed on base and extremely limited road access between the base and the closest urban center. The relative lack of venues where military men may meet sexual partners due to the remote nature of the base may have limited sexual networking between the local military population on Base 2 and AGYW. In addition, there may not be enough HIV-positive military men in comparison to the general population of men to contribute significantly to the local HIV epidemic.

There are a number of features that distinguish Base 3 from the other military bases in our study. AGYW at Base 3 had the highest rates of new HIV-positive diagnoses. In addition, AGYW proximal to Base 3 were the youngest and least educated of the four military bases. In addition, substantially higher proportions of AGYW reported transactional sex (13.2%) compared to the next highest proportion at Base 1 (8.4%). Base 3 is located a few kilometers outside an urban center; however, there is much better road connectivity between the urban center and Base 3's location. AGYW were recruited across the entire north-south corridor below Base 3, providing substantial heterogeneity in the locations where AGYW

were recruited. Our findings clearly show that AGYW that were recruited from proximal locations to Base 3 were at substantially higher risk of HIV-infection compared to AGYW with less access. Consequently, our findings may highlight an important local sexual network that could be targeted for additional HIV prevention, care, and treatment interventions.

There are a number of important strengths in our research. There has been no published HIV/AIDS research on the interaction between local military populations and their surrounding communities in any Sub-Saharan African military prior to this study. Current applications of geospatial analysis to epidemiologic studies in Mozambique have been limited.[121–123] In addition, there have been no previous studies applying geospatial analysis to Sub-Saharan African military populations. The precision and scale of the geographic information used in this study is unique to HIV/AIDS research: sub-Saharan African military populations, sub-Saharan African adolescent girls and young women, and these populations specifically in Mozambique. Finally, the construction of our measure of geographic access to specific locations was able to account for important features that determine how people actually move between two locations.

Our study was unable to publish extensive information on the geographic characteristics of the military bases and military men stationed at these locations due to security sensitivities. However, we believe that the features of these bases that have been identified in this paper are descriptive enough to provide readers with a fair sense of the key differences between these military bases and the communities in which they are co-located. We are unable to directly assess sexual relationships between military personnel and AGYW in surrounding communities due to women not being aware that their partner is a member of the military (e.g. plainclothes) or hesitance to report military partners due to gender-power dynamics. Instead, our study hypothesized that proximity to military facilities is an indicator of accessibility to military sexual partners. While we observed a small, significant effect supporting this hypothesis in our overall sample, our stratified results clearly show that this hypothesis does not hold true for all military bases. Features that distinguish Base 3 from the other bases may highlight important

features that can inform improved targeting of HIV-resources to military locations where they may be more effective than others.

It is possible that the sampling of young women from locations where they congregate or meet sexual partners may not be representative of the entire population of women in this age group in these communities. It is not feasible to conduct random household sampling given that many of these women are not present in their homes during the day as they attend to their daily tasks attending schools, buying or selling in markets, or other activities. Conducting study activities outside of daylight hours is often unsafe for study staff, which limits the times at which study sampling can be completed. However, our survey results indicate wide geographic variation in where the study population of young women have been recruited and geographic variation in locations where HIV-positive women have been found.

Our results indicate that there may be important local sexual networks between AGYW congregating or meeting partners in communities surrounding 1 of our 4 military bases. As HIV prevention, care, and treatment programs are being asked to increase their impact without subsequent increases in their financial resources, improved targeting of these interventions will be important to achieve UNAIDS 90-90-90 goals. Our study provides call for a more detailed examination and potential targeting of HIV resources in one of the military bases included in our study. In addition, military facilities that have similar features to this may represent better targets for HIV resources than military facilities that more closely the features of the remaining three bases. While we have used geospatial methods as an indicator of sexual networking between these two important populations, future research stemming from this study will use HIV-1 pol sequence analysis to better characterize the intermixing of military and civilian HIV epidemics at these four bases.

Chapter 4, in full, is currently being prepared for submission and publication of the material. Davitte, J.; Pines, H.; Martin, N.; Brodine, S.; Salem, R.; Shaffer, R. The dissertation author was the primary investigator and author of this material.

Table 8: Estimates of travel time by road and surface type

Surface	**Travel Speed**[a] (kilometers/hour)	**Time**[b] (minutes/5 meters)
Paved Major Road	80	0.00375
Unpaved Major Road	60	0.005
Paved Minor Road	4	0.075
Unpaved Minor Road	4	0.075
Trail	3	0.1
General Terrain	3	0.1
Water	Impassable (99,999 minutes/10m)	

[a] *Travel speed defined according to methodology described by Makanga, et.al*
[b] *Time in minutes required to cross 5 meters for a specific road and surface type*

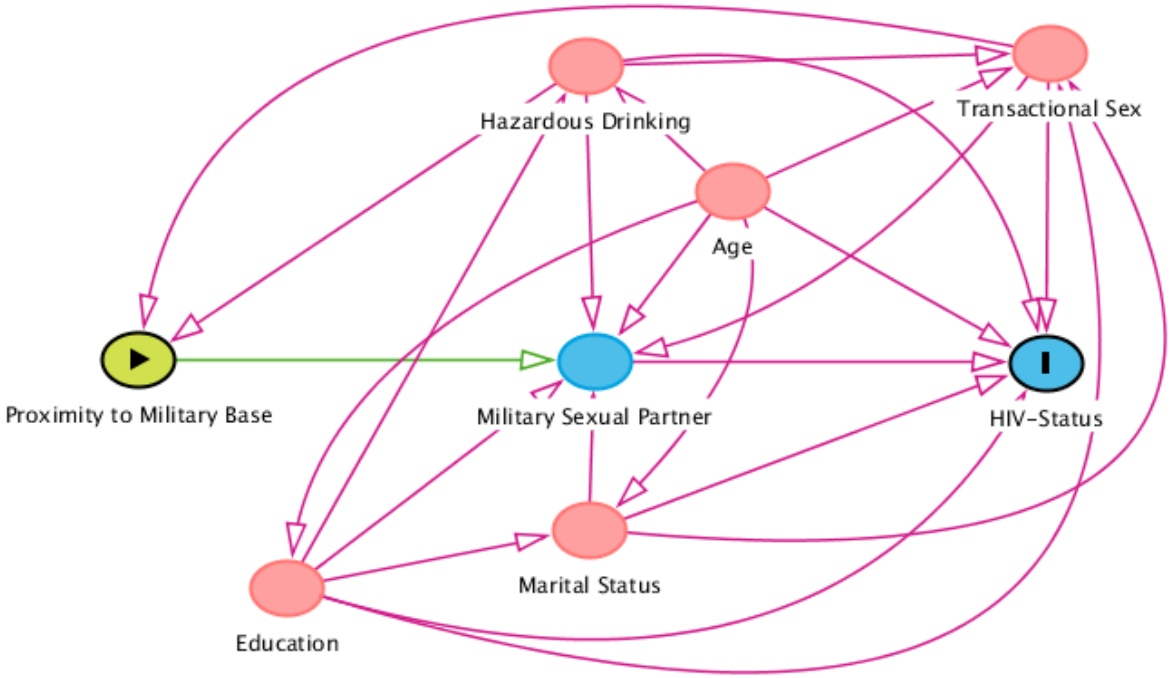


Figure 6: Directed acyclic graph for association between proximity to military bases for locations where adolescent girls and young women congregate or meet their sexual partners and HIV status

Table 9: Participant characteristics by nearest military base, Mozambique, 2018-2019 (n=7,514)

Variable	Base 1 *(N=2772)*	Base 2 *(N=1759)*	Base 3 *(N=1386)*	Base 4 *(N=1597)*	Total *(N=7514)*
HIV-positive status, n (%)	161 (5.8%)	88 (5.0%)	88 (6.3%)	136 (8.5%)	473 (6.3%)
Newly diagnosed with HIV[a], Yes, n (%)	115 (4.2%)	64 (3.7%)	71 (5.2%)	71 (4.6%)	321 (4.4%)
Travel Time[b], minutes	45.5 (22.6)	83.5 (26.2)	57.1 (15.6)	30.3 (17.6)	53.3 (28.5)
Age, years, mean (SD)	23.2 (5.4)	24.2 (6.0)	22.5 (5.1)	23.5 (5.3)	23.4 (5.5)
Current marital status, n (%)					
Single, never married, and not living with a partner	1329 (48.0%)	843 (47.9%)	579 (41.8%)	424 (26.5%)	3175 (42.3%)
Single, living with a partner	416 (15.0%)	260 (14.8%)	187 (13.5%)	557 (34.9%)	1420 (18.9%)
Married or in union	749 (27.0%)	474 (26.9%)	588 (42.4%)	472 (29.6%)	2283 (30.4%)
Other	277 (10.0%)	182 (10.3%)	32 (2.3%)	144 (9.0%)	635 (8.5%)
Highest Education Achieved, n (%)					
Some Primary or None	497 (18.6%)	399 (23.5%)	369 (26.8%)	149 (10.2%)	1414 (19.6%)
Primary	1529 (57.3%)	1043 (61.5%)	771 (56.1%)	820 (56.2%)	4163 (57.9%)
Secondary or Higher	641 (24.0%)	254 (15.0%)	235 (17.1%)	489 (33.5%)	1619 (22.5%)
Military Sexual Partner[c], Yes, n (%)	55 (2.0%)	83 (4.7%)	55 (4.0%)	103 (6.4%)	296 (3.9%)
Hazardous drinking[d], Yes, n (%)	478 (17.2%)	263 (15.0%)	66 (4.8%)	55 (3.4%)	862 (11.5%)
Transactional sex[e], Yes, n (%)	233 (8.4%)	64 (3.6%)	183 (13.2%)	93 (5.8%)	573 (7.6%)

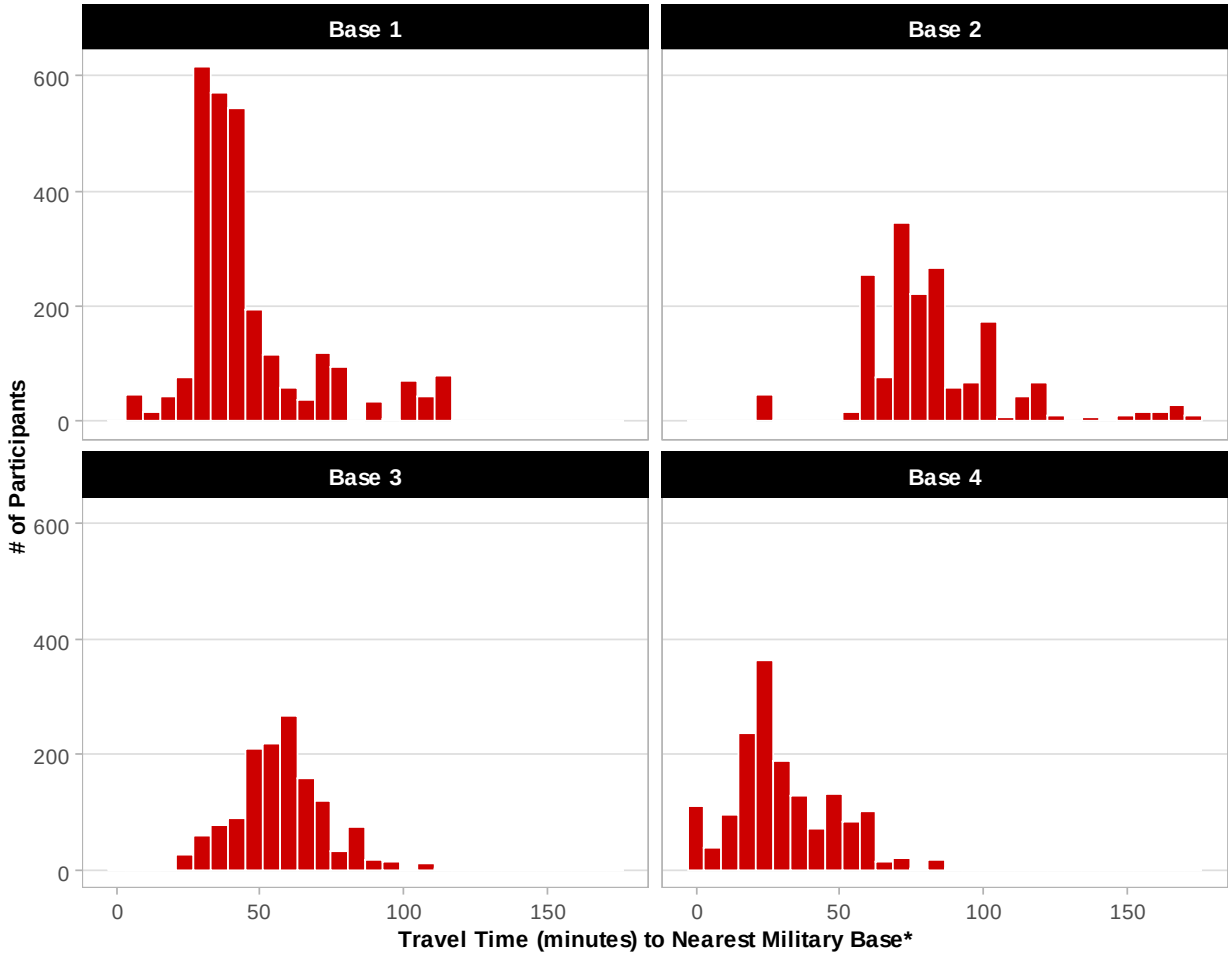
[a] *Among HIV-positive participants, those that did not indicate "HIV-positive" status at time of cross-sectional survey*

[b] *Estimated time in minutes it would take a female participant to travel (walking + public transportation) from their recruitment location to the closet military base (accounting for terrain, road, and road surface type)*

[c] *Participant stated that 1 of their 3 most recent sexual partners was a member of the military*

[d] *Measured by an Alcohol Use Disorders Identification Test (AUDIT-C) score of 3 or greater*

[e] *Received money, shelter, food, drugs, favors, or gifts in exchange for sex with any of three most recent sexual partners*



** Estimated time in minutes it would take a female participant to travel (walking + public transportation) from their recruitment location to the closet military base (accounting for terrain, road, and road surface type)*

Figure 7: Estimated Travel Time (minutes) from Participant Recruitment location to Closest Military Base, Mozambique, 2018-2019 (n=7,514)

Table 10: Participant characteristics by current HIV-status and nearest military base (Base 1 & 2), Mozambique, 2018-2019 (n=7,514)

Variable	Base 1		Base 2	
	HIV-Negative *(n=2611)*	**HIV-Positive** *(n=161)*	**HIV-Negative** *(n=1671)*	**HIV-Positive** *(n=88)*
Travel Time[a], minutes	45.5 (22.4)	45.3 (24.9)	83.4 (26.3)	84.9 (24.7)
Age, years, mean (SD)	23.0 (5.3)	26.6 (5.4)	24.1 (5.9)	27.1 (5.8)
Current marital status, n (%)				
Single, never married, and not living with a partner	1308 (50.1%)	21 (13.0%)	833 (49.9%)	10 (11.4%)
Single, living with a partner	386 (14.8%)	30 (18.6%)	238 (14.2%)	22 (25.0%)
Married or in union	679 (26.0%)	70 (43.5%)	442 (26.5%)	32 (36.4%)
Other	237 (9.1%)	40 (24.8%)	158 (9.5%)	24 (27.3%)
Highest Education Achieved, n (%)				
Some Primary or None	466 (18.5%)	31 (20.7%)	362 (22.5%)	37 (43.5%)
Primary	1454 (57.8%)	75 (50.0%)	1008 (62.6%)	35 (41.2%)
Secondary or Higher	597 (23.7%)	44 (29.3%)	241 (15.0%)	13 (15.3%)
Hazardous drinking[b], Yes, n (%)	432 (16.5%)	46 (28.6%)	227 (13.6%)	36 (40.9%)
Transactional sex[c], Yes, n (%)	214 (8.2%)	19 (11.8%)	54 (3.2%)	10 (11.4%)
Military Sexual Partner[d], Yes, n (%)	52 (2.0%)	3 (1.9%)	81 (4.8%)	2 (2.3%)
Newly diagnosed with HIVa, Yes, n (%)	0 (0.0%)	115 (71.4%)	0 (0.0%)	64 (72.7%)

[a] *Estimated time in minutes it would take a female participant to travel (walking + public transportation) from their recruitment location to the closet military base (accounting for terrain, road, and road surface type)*

[b] *Measured by an Alcohol Use Disorders Identification Test (AUDIT-C) score of 3 or greater*

[c] *Received money, shelter, food, drugs, favors, or gifts in exchange for sex with any of three most recent sexual partners*

[d] *Participant stated that 1 of their 3 most recent sexual partners was a member of the military*

[e] *Among HIV-positive participants, those that did not indicate "HIV-positive" status at time of cross-sectional survey*

Table 11: Participant characteristics by current HIV-status and nearest military base (Base 3 & 4), Mozambique, 2018-2019 (n=7,514)

Variable	Base 3		Base 4	
	HIV-Negative *(n=2611)*	**HIV-Positive** *(n=161)*	**HIV-Negative** *(n=1671)*	**HIV-Positive** *(n=88)*
Travel Time[a], minutes	57.4 (15.6)	52.6 (15.7)	30.6 (17.5)	26.9 (18.6)
Age, years, mean (SD)	22.3 (5.0)	25.3 (5.4)	23.1 (5.2)	27.1 (5.3)
Current marital status, n (%)				
Single, never married, and not living with a partner	557 (42.9%)	22 (25.0%)	408 (27.9%)	16 (11.8%)
Single, living with a partner	165 (12.7%)	22 (25.0%)	518 (35.5%)	39 (28.7%)
Married or in union	549 (42.3%)	39 (44.3%)	425 (29.1%)	47 (34.6%)
Other	27 (2.1%)	5 (5.7%)	110 (7.5%)	34 (25.0%)
Highest Education Achieved, n (%)				
Some Primary or None	353 (27.4%)	16 (18.4%)	127 (9.6%)	22 (17.1%)
Primary	723 (56.1%)	48 (55.2%)	765 (57.6%)	55 (42.6%)
Secondary or Higher	212 (16.5%)	23 (26.4%)	437 (32.9%)	52 (40.3%)
Hazardous drinking[b], Yes, n (%)	59 (4.5%)	7 (8.0%)	47 (3.2%)	8 (5.9%)
Transactional sex[c], Yes, n (%)	155 (11.9%)	28 (31.8%)	89 (6.1%)	4 (2.9%)
Military Sexual Partner[d], Yes, n (%)	47 (3.6%)	8 (9.1%)	89 (6.1%)	14 (10.3%)
Newly diagnosed with HIV[e], Yes, n (%)	0 (0.0%)	71 (44.1%)	0 (0.0%)	71 (80.7%)

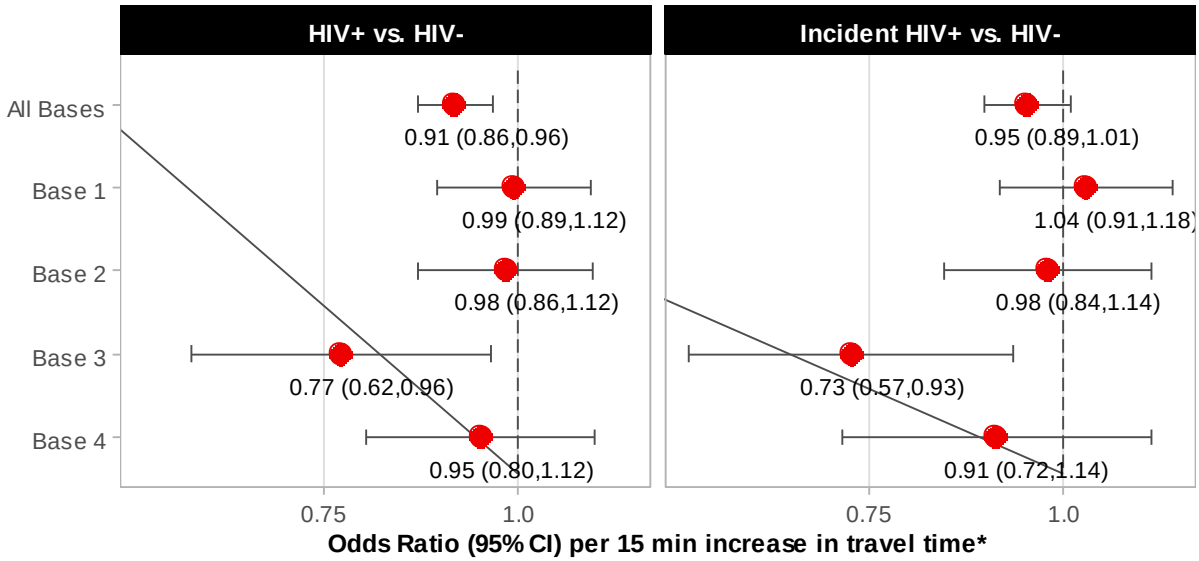
[a] *Estimated time in minutes it would take a female participant to travel (walking + public transportation) from their recruitment location to the closet military base (accounting for terrain, road, and road surface type)*

[b] *Measured by an Alcohol Use Disorders Identification Test (AUDIT-C) score of 3 or greater*

[c] *Received money, shelter, food, drugs, favors, or gifts in exchange for sex with any of three most recent sexual partners*

[d] *Participant stated that 1 of their 3 most recent sexual partners was a member of the military*

[e] *Among HIV-positive participants, those that did not indicate "HIV-positive" status at time of cross-sectional survey*



** Distance in kilometers of the most optimal travel route (accounting for terrain, road, and road surface type) between a female participant's recruitment location and the closest military base*
*** HIV+ vs. HIV-: comparison of all HIV-positive participants to all HIV-negative participant; Incident HIV+ vs. HIV-: comparison of newly-diagnosed HIV-positive participants to all HIV-negative participants*
**** Multivariable logistic regression, adjusting for participant age, hazardous drinking, marital status, transactional sex, and education*

Figure 8: Adjusted odds of HIV-positive status per in travel time on optimal route from recruitment location to nearest military base, Mozambique, 2018-2019 (n=7,514)

Chapter 5: Discussion

The COVID-19 epidemic has highlighted key challenges for infectious disease Epidemiologic research: scaling causal inference efforts across the human disease phenome, understanding the long-term consequences of a novel disease without robust longitudinal data, and leveraging non-traditional types of data for infectious disease research. Our dissertation has provided three examples of advanced Epidemiologic methods that illustrate how researchers may address one or more of these challenges.

In Chapter 2, we focused on addressing the issue of scalability. We identified 5 distinct comorbidity patterns from 31 disease indicators, assessed using clinical diagnosis records from UK Biobank's comprehensive EHR data linkage between 2015-2019. Our results identified significantly increased risk for severe COVID-19 infection for our 'Some Non-Specific Health Conditions', 'Diabetics with 1-2 other conditions', 'Cardiac multimorbidity', and 'Cancer multimorbidity' latent classes compared to our 'Healthy' latent class. In addition, our results identified substantial heterogeneity in the effect sizes of severe COVID-19 infection risk between our comorbidity latent classes. Our use of a 3-step, bias-adjusted LCA with a distal outcome, provides a concrete example of an alternative method to traditional 'simultaneous estimate' and 'disease-by-disease' approaches for investigating the relationships between a large number of important exposures and an important disease outcome. In stark contrast to the '*simultaneous estimate*' and '*disease-by-disease*' approaches, our LCA application was not only able to identify important disease patterns from EHR data, but it was also able to robustly measure the association between those patterns and a novel disease outcome.

In Chapter 3, we focused on the challenges of understanding novel disease long-term consequences and leveraging non-traditional data types. We investigated the association between genetic liability to severe COVID-19 infection, measured via PRSs, and 31 comorbidity phenotypes derived from linked EHR data collected over the past 20 years. We identified shared genetic etiology between severe COVID-19 infection and 4 comorbidities ('Diabetes, uncomplicated', 'Hypertension, uncomplicated', 'Obesity', and 'Renal Failure') in the UK Biobank cohort, representing the most comprehensive

assessment of shared genetic risk for severe COVID-19 and EHR-derived comorbidities to date. Our research indicates that the same genetic composition that increases an individual's risk for COVID-19 may also influence their risk for other important comorbid diseases. This example demonstrates how the combination of genetic data and EHR data can successfully inform future research on consequences of a novel disease, through the identification of shared genetic risk, even in the absence of accumulated longitudinal data post-disease emergence.

Finally, in Chapter 4, we again focused on the challenge of leveraging non-traditional data types to inform infectious disease research. Our study leveraged GIS data and methods to construct a novel exposure (proximity to military bases, accounting for important features that determine how people actually move between two locations) to study the interactions between two populations with known risk for HIV infection. Our study found that AGYW that meet or congregate near military bases were at a slightly elevated risk for HIV-infection in the combined sample. In our stratified analysis by military base, we observed a strong relationship between HIV-positive diagnosis (both overall and in the subset of newly diagnosed AGYW) in only 1 of our 4 military bases. While our findings supported the hypothesis that AGYW congregating or meeting sexual partners at venues in closer proximity to military bases is positively associated with HIV infection in our overall sample, our stratified analysis indicated that this hypothesis did not hold true across all types of military bases and the communities that surround them. This example shows how the use of GIS data can support research in situations where sensitivities around infectious disease prevention may inhibit direct questioning or contact tracing.

One common limitation across each of the examples in our dissertation is that we did not measure causal relationships between our exposures and our outcomes of interests. Causal inference research typically requires pre-existing knowledge (often derived from prior research) of the potential confounders that may impact exposure-outcome effect estimates in order to accurately quantify the exposure-outcome relationship. In the context of our COVID-19 research, there has not been sufficient data nor time to appropriately inform causal models across the entire disease phenome. In the context of our HIV/AIDS research, we are unable to directly assess a causal relationship between military sexual partners and HIV

infection in AGYW due to difficulty in directly assesses sexual relationships between these populations. In all three of the examples, the intention of our research was to improve targeting of future causal-inference focused research.

An additional limitation specific to the examples in Chapters 2 and 3 is that comorbidity measures were derived exclusively to diagnosis codes, leading to potential for within-disease heterogeneity due to coding practices by different healthcare providers; as well as vulnerabilities due to provider ascertainment of sufficient disease to warrant the recording of a diagnosis code. In addition, the examples in Chapters 2 and 3 were unable to leverage diagnosis data stored in EMIS practices, which reduced sample sizes in both research efforts. An important limitation specific to our Chapter 4 research was the possibility that our sampling of AGYW from locations where they congregate or meet sexual partners may not have been representative of the entire population of women in this age group in these communities.

There are a number of notable strengths to our dissertation research. COVID-19 studies that leverage UK Biobank data frequently use self-reported conditions, biomarkers, and other variables that were measured during the Baseline Assessment visit between 2006 and 2010. Each of our comorbidity measures in Chapters 2 and 3 were developed through diagnosis codes, which required health care providers to record the diagnosis based on a clinical evaluation. In addition, each example in Chapters 2 and 3 were able to leverage the UK Biobank's primary care data linkage, improving our ability to completely ascertain health status of UK Biobank participants seen within TPP GP practices. Furthermore, for both Chapters 2 and 3, we did not select comorbidities for investigation based on our own opinions, information form prior research, or with the intention to validate a specific hypothesis. Rather, we used a common measure of comorbidities that is used across many research settings. Strengths of our Chapter 4 example are a bit more specific to the research question that was investigated. There have been no published HIV/AIDS research on the interaction between local military populations and their surrounding communities in any Sub-Saharan African military prior to this study. In addition, there have been no previous studies applying geospatial analysis to Sub-Saharan African military

populations. Furthermore, the precision and scale of the geographic information used in this study is unique to HIV/AIDS research: sub-Saharan African military populations, sub-Saharan African adolescent girls and young women, and these populations specifically in Mozambique.

In this dissertation, we have provided real-world examples of investigating infectious disease that leverage numerous types of health data (genetic, survey, EHR, and registry data), use distinct Epidemiologic methods (LCA, PRS, and GIS), and address specific challenges in researching a novel disease such as COVID-19 (scalability, long-term consequences, and integration of non-traditional data types). Our Chapter 2 example demonstrates how the use of LCA with a distal outcome can robustly highlight disease patterns that are relevant for an outcome of interest without significant, *a priori* knowledge of the relationships between the diseases. Our Chapter 3 example demonstrates how the combination of genetic data and EHR data can successfully inform researchers on potential future consequences of a novel disease, through the identification of shared genetic risk, even in the absence of accumulated longitudinal data post-disease emergence. Finally, our Chapter 4 example demonstrates how GIS data can be used to ascertain the relationship between a proxy for exposure and a specific infectious disease outcome in situations where directly measuring the exposure is not feasible. These examples provide a potential roadmap for how Epidemiologists may respond to novel or emerging infectious disease epidemics in the future.

References

1. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. *Annu Rev Public Health*. 2016;37:61-81. doi:10.1146/annurev-publhealth-032315-021353

2. Galea S, Tracy M. Participation Rates in Epidemiologic Studies. *Ann Epidemiol*. 2007;17(9):643-653. doi:10.1016/j.annepidem.2007.03.013

3. Denaxas S, Gonzalez-Izquierdo A, Direk K, et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. *J Am Med Informatics Assoc*. 2019;26(12):1545-1559. doi:10.1093/jamia/ocz105

4. Turnbull C, Scott RH, Thomas E, et al. The 100 000 Genomes Project: Bringing whole genome sequencing to the NHS. *BMJ*. 2018;361. doi:10.1136/bmj.k1687

5. 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(3). doi:10.1371/journal.pmed.1001779

6. UK Biobank. COVID-19 data. Published 2020. Accessed January 21, 2021. https://www.ukbiobank.ac.uk/enable-your-research/about-our-data/covid-19-data

7. Krieger N. Place, space, and health: GIS and epidemiology. *Epidemiology*. 2003;14(4):384-385. doi:10.1097/01.ede.0000071473.69307.8a

8. Smith CD, Mennis J. Incorporating geographic information science and technology in response to the COVID-19 pandemic. *Prev Chronic Dis*. 2020;17(17). doi:10.5888/PCD17.200246

9. Cao P, Song Y, Zhuang Z, et al. Obesity and COVID-19 in Adult Patients With Diabetes. *Diabetes*. Published online February 17, 2021:db200671. doi:10.2337/db20-0671

10. Gao C, Gao C, Cai Y, et al. Association of hypertension and antihypertensive treatment with COVID-19 mortality: a retrospective observational study. *Eur Heart J*. 2020;41(22):2058-2066. doi:10.1093/eurheartj/ehaa433

11. Velásquez-Tirado JD, Trzepacz PT, Franco JG. Etiologies of Delirium in Consecutive COVID-19 Inpatients and the Relationship Between Severity of Delirium and COVID-19 in a Prospective Study With Follow-Up. *J Neuropsychiatry Clin Neurosci*. Published online April 12, 2021. doi:10.1176/appi.neuropsych.20100251

12. Tartof SY, Qian L, Hong V, et al. Obesity and Mortality Among Patients Diagnosed With COVID-19: Results From an Integrated Health Care Organization. *Ann Intern Med*. 2020;173(10):773-781. doi:10.7326/M20-3742

13. Panagiotou OA, Kosar CM, White EM, et al. Risk Factors Associated with All-Cause 30-Day Mortality in Nursing Home Residents with COVID-19. *JAMA Intern Med*. 2021;181(4):439-448. doi:10.1001/jamainternmed.2020.7968

14. Ioannou GN, Locke E, Green P, et al. Risk Factors for Hospitalization, Mechanical Ventilation, or

Death Among 10 131 US Veterans With SARS-CoV-2 Infection. *JAMA Netw open*. 2020;3(9):e2022310. doi:10.1001/jamanetworkopen.2020.22310

15. Westreich D, Greenland S. The table 2 fallacy: Presenting and interpreting confounder and modifier coefficients. *Am J Epidemiol*. 2013;177(4):292-298. doi:10.1093/aje/kws412

16. Vermunt JK, Magidson J. How to perform three-step latent class analysis in the presence of measurement non-invariance or differential item functioning. *Struct Equ Model A Multidiscip J*. Published online September 22, 2020:1-9. doi:10.1080/10705511.2020.1818084

17. Renu K, Prasanna PL, Valsala Gopalakrishnan A. Coronaviruses pathogenesis, comorbidities and multi-organ damage – A review. *Life Sci*. 2020;255. doi:10.1016/j.lfs.2020.117839

18. Pairo-Castineira E, Clohisey S, Klaric L, et al. Genetic mechanisms of critical illness in Covid-19. *Nature*. Published online December 11, 2020. doi:10.1038/s41586-020-03065-y

19. Visscher PM, Wray NR, Zhang Q, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet*. Published online 2017. doi:10.1016/j.ajhg.2017.06.005

20. Richardson TG, Harrison S, Hemani G, Smith GD. An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome. *Elife*. Published online 2019. doi:10.7554/eLife.43657

21. Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev*. 2009;19(3):212-219. doi:10.1016/j.gde.2009.04.010

22. Sugrue LP, Desikan RS. What Are Polygenic Scores and Why Are They Important? *JAMA - J Am Med Assoc*. 2019;321(18):1820-1821. doi:10.1001/jama.2019.3893

23. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet*. Published online 2018. doi:10.1038/s41576-018-0018-x

24. Choi SW, Mak TSH, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc*. 2020;15(9):2759-2772. doi:10.1038/s41596-020-0353-1

25. Xicoy H, Klemann CJ, De Witte W, Martens MB, Martens GJ, Poelmans G. Shared genetic etiology between Parkinson's disease and blood levels of specific lipids. *npj Park Dis*. 2021;7(1):1-8. doi:10.1038/s41531-021-00168-9

26. Andersen AM, Pietrzak RH, Kranzler HR, et al. Polygenic Scores for Major Depressive Disorder and Risk of Alcohol Dependence. *JAMA psychiatry*. 2017;74(11):1153-1160. doi:10.1001/jamapsychiatry.2017.2269

27. Takahashi N, Nishimura T, Harada T, et al. Polygenic risk score analysis revealed shared genetic background in attention deficit hyperactivity disorder and narcolepsy. *Transl Psychiatry*. 2020;10(1):1-9. doi:10.1038/s41398-020-00971-7

28. Foo JC, Streit F, Treutlein J, et al. Shared genetic etiology between alcohol dependence and major depressive disorder. *Psychiatr Genet*. 2018;28(4):66-70. doi:10.1097/YPG.0000000000000201

29. Lutz MW, Sprague D, Barrera J, Chiba-Falek O. Shared genetic etiology underlying Alzheimer's disease and major depressive disorder. *Transl Psychiatry*. 2020;10(1):1-14. doi:10.1038/s41398-020-0769-y

30. Hohman M, McMaster F, Woodruff SI. Contact Tracing for COVID-19: The Use of Motivational Interviewing and the Role of Social Work. *Clin Soc Work J*. 2021;1:1. doi:10.1007/s10615-021-00802-2

31. Johns Hopkins Coronavirus Resource Center. COVID-19 Map - Johns Hopkins Coronavirus Resource Center. Accessed February 16, 2021. https://coronavirus.jhu.edu/map.html

32. Johns Hopkins Coronavirus Resource Center. Mortality Analyses - Johns Hopkins Coronavirus Resource Center. Accessed February 16, 2021. https://coronavirus.jhu.edu/data/mortality

33. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. *Med Care*. 1998;36(1):8-27. doi:10.1097/00005650-199801000-00004

34. Møller SP, Laursen B, Johannesen CK, Tolstrup JS, Schramm S. Patterns of multimorbidity and demographic profile of latent classes in a Danish population—A register-based study. Devleesschauwer B, ed. *PLoS One*. 2020;15(8):e0237375. doi:10.1371/journal.pone.0237375

35. 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. *Lancet*. Published online 2012. doi:10.1016/S0140-6736(12)60240-2

36. Marengoni A, Angleman S, Melis R, et al. Aging with multimorbidity: A systematic review of the literature. *Ageing Res Rev*. Published online 2011. doi:10.1016/j.arr.2011.03.003

37. Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care*. Published online 2005. doi:10.1097/01.mlr.0000182534.19832.83

38. 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*. Published online 2019. doi:10.1186/s12874-019-0753-5

39. Veenendaal M, Westerik JAM, van den Bemt L, Kocks JWH, Bischoff EW, Schermer TR. Age- and sex-specific prevalence of chronic comorbidity in adult patients with asthma: A real-life study. *npj Prim Care Respir Med*. 2019;29(1):1-7. doi:10.1038/s41533-019-0127-9

40. Ahrenfeldt LJ, Möller S, Thinggaard M, Christensen K, Lindahl-Jacobsen R. Sex Differences in Comorbidity and Frailty in Europe. *Int J Public Health*. 2019;64(7):1025-1036. doi:10.1007/s00038-019-01270-9

41. Abad-Díez JM, Calderón-Larrañaga A, Poncel-Falcó A, et al. Age and gender differences in the prevalence and patterns of multimorbidity in the older population. *BMC Geriatr*. 2014;14(1):75. doi:10.1186/1471-2318-14-75

42. R: The R Project for Statistical Computing. Accessed January 21, 2021. https://www.r-project.org/

43. Latent GOLD® 5.1 - Statistical Innovations. Accessed January 21, 2021. https://www.statisticalinnovations.com/latent-gold-5-1/

44. Ulbricht CM, Chrysanthopoulou SA, Levin L, Lapane KL. The use of latent class analysis for identifying subtypes of depression: A systematic review. *Psychiatry Res*. 2018;266:228-246. doi:10.1016/j.psychres.2018.03.003

45. Collins LM, Lanza ST. *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*. John Wiley and Sons Inc.; 2010. doi:10.1002/9780470567333

46. Park HC, Kim DH, Cho A, et al. Clinical outcomes of initially asymptomatic patients with COVID-19: a Korean nationwide cohort study. *Ann Med*. 2021;53(1):357-364. doi:10.1080/07853890.2021.1884744

47. Fagard K, Gielen E, Deschodt M, Devriendt E, Flamaing J. Risk factors for severe COVID-19 disease and death in patients aged 70 and over: a retrospective observational cohort study. *Acta Clin Belg*. Published online February 21, 2021:1-8. doi:10.1080/17843286.2021.1890452

48. Mak JKL, Kuja-Halkola R, Wang Y, Hägg S, Jylhävä J. Frailty and comorbidity in predicting community <scp>COVID</scp> -19 mortality in the <scp>UK</scp> Biobank: the effect of sampling. *J Am Geriatr Soc*. Published online February 22, 2021:jgs.17089. doi:10.1111/jgs.17089

49. Imam Z, Odish F, Gill I, et al. Older age and comorbidity are independent mortality predictors in a large cohort of 1305 COVID-19 patients in Michigan, United States. *J Intern Med*. 2020;288(4):469-476. doi:10.1111/joim.13119

50. Williamson EJ, Walker AJ, Bhaskaran K, et al. Factors associated with COVID-19-related death using OpenSAFELY. *Nature*. 2020;584(7821):430-436. doi:10.1038/s41586-020-2521-4

51. Westreich D, Edwards JK, van Smeden M. Comment on Williamson et al. (OpenSAFELY): The Table 2 Fallacy in a Study of COVID-19 Mortality Risk Factors. *Epidemiology*. 2021;32(1):e1-e2. doi:10.1097/EDE.0000000000001259

52. Kumar A, Arora A, Sharma P, et al. Is diabetes mellitus associated with mortality and severity of COVID-19? A meta-analysis. *Diabetes Metab Syndr Clin Res Rev*. 2020;14(4):535-545. doi:10.1016/j.dsx.2020.04.044

53. Wang W, Shen M, Tao Y, et al. Elevated glucose level leads to rapid COVID-19 progression and high fatality. *BMC Pulm Med*. 2021;21(1):64. doi:10.1186/s12890-021-01413-w

54. Li H, Tian S, Chen T, et al. Newly diagnosed diabetes is associated with a higher risk of mortality than known diabetes in hospitalized patients with <scp>COVID</scp> -19. *Diabetes, Obes Metab*. 2020;22(10):1897-1906. doi:10.1111/dom.14099

55. Li P, Wu W, Zhang T, et al. Implications of cardiac markers in risk-stratification and management for COVID-19 patients. *Crit Care*. 2021;25(1):158. doi:10.1186/s13054-021-03555-z

56. Li X, Zhong X, Wang Y, Zeng X, Luo T, Liu Q. Clinical determinants of the severity of COVID-19: A systematic review and meta-analysis. *PLoS One*. 2021;16(5):e0250602. doi:10.1371/journal.pone.0250602

57. Wingert A, Pillay J, Gates M, et al. Risk factors for severity of COVID-19: a rapid review to inform vaccine prioritisation in Canada. *BMJ Open*. 2021;11(5):e044684. doi:10.1136/bmjopen-2020-044684

58. Krasnow MR, Litt HK, Lehmann CJ, Lio J, Zhu M, Sherer R. Cancer, transplant, and immunocompromising conditions were not significantly associated with severe illness or death in

hospitalized COVID-19 patients. *J Clin Virol*. 2021;140:104850. doi:10.1016/j.jcv.2021.104850

59. Meng Y, Meng Y, Lu W, et al. Cancer history is an independent risk factor for mortality in hospitalized COVID-19 patients: A propensity score-matched analysis. *J Hematol Oncol*. 2020;13(1):1-11. doi:10.1186/s13045-020-00907-0

60. Xia Y, Jin R, Zhao J, Li W, Shen H. Risk of COVID-19 for patients with cancer. *Lancet Oncol*. 2020;21(4):e180. doi:10.1016/S1470-2045(20)30150-9

61. Liang W, Guan W, Chen R, et al. Cancer patients in SARS-CoV-2 infection: a nationwide analysis in China. *Lancet Oncol*. 2020;21(3):335-337. doi:10.1016/S1470-2045(20)30096-6

62. Grint DJ, Wing K, Williamson E, et al. Case fatality risk of the SARS-CoV-2 variant of concern B.1.1.7 in England, 16 November to 5 February. *Euro Surveill*. 2021;26(11). doi:10.2807/1560-7917.ES.2021.26.11.2100256

63. Davies NG, Jarvis CI, van Zandvoort K, et al. Increased mortality in community-tested cases of SARS-CoV-2 lineage B.1.1.7. *Nature*. Published online 2021. doi:10.1038/s41586-021-03426-1

64. Challen R, Brooks-Pollock E, Read JM, Dyson L, Tsaneva-Atanasova K, Danon L. Risk of mortality in patients infected with SARS-CoV-2 variant of concern 202012/1: Matched cohort study. *BMJ*. 2021;372. doi:10.1136/bmj.n579

65. Davies NG, Abbott S, Barnard RC, et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science (80-)*. 2021;372(6538):eabg3055. doi:10.1126/science.abg3055

66. Davies NG, Jarvis CI, van Zandvoort K, et al. Increased mortality in community-tested cases of SARS-CoV-2 lineage B.1.1.7. *Nature*. Published online March 15, 2021:1-5. doi:10.1038/s41586-021-03426-1

67. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *Eur J Hum Genet*. 2020;28(6):715-718. doi:10.1038/s41431-020-0636-6

68. Liu D, Yang J, Feng B, Lu W, Zhao C, Li L. Mendelian randomization analysis identified genes pleiotropically associated with the risk and prognosis of COVID-19. *medRxiv*. Published online September 4, 2020. doi:10.1101/2020.09.02.20187179

69. COVID-19 Host Genetics Initiative. COVID-19 HGI Results for Data Freeze 4 (October 2020) | Blog. Accessed February 16, 2021. https://www.covid19hg.org/blog/2020-11-24-covid-19-hgi-results-for-data-freeze-4-october-2020/

70. Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience*. Published online 2019. doi:10.1093/gigascience/giz082

71. Coombes BJ, Ploner A, Bergen SE, Biernacka JM. A principal component approach to improve association testing with polygenic risk scores. *Genet Epidemiol*. 2020;44(7):676-686. doi:10.1002/gepi.22339

72. Ganna A. Mapping the human genetic architecture of COVID-19 by worldwide meta-analysis. *medRxiv*. Published online March 12, 2021:2021.03.10.21252820. doi:10.1101/2021.03.10.21252820

73. Shah S, Henry A, Roselli C, et al. Genome-wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure. *Nat Commun*. 2020;11(1). doi:10.1038/s41467-019-13690-5

74. Wuttke M, Li Y, Li M, et al. A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat Genet*. 2019;51(6):957-972. doi:10.1038/s41588-019-0407-x

75. Caussy C, Wallet F, Laville M, Disse E. Obesity is Associated with Severe Forms of COVID-19. *Obesity*. 2020;28(7):1175. doi:10.1002/oby.22842

76. Simonnet A, Chetboun M, Poissy J, et al. High Prevalence of Obesity in Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) Requiring Invasive Mechanical Ventilation. *Obesity*. 2020;28(7):1195-1199. doi:10.1002/oby.22831

77. Svensson P, Hofmann R, Häbel H, Jernberg T, Nordberg P. Association between cardiometabolic disease and severe COVID-19: a nationwide case–control study of patients requiring invasive mechanical ventilation. *BMJ Open*. 2021;11(2):e044486. doi:10.1136/bmjopen-2020-044486

78. Aung N, Khanji MY, Munroe PB, Petersen SE. Causal Inference for Genetic Obesity, Cardiometabolic Profile and COVID-19 Susceptibility: A Mendelian Randomization Study. *Front Genet*. 2020;11:1417. doi:10.3389/fgene.2020.586308

79. Stefan N, Birkenfeld AL, Schulze MB, Ludwig DS. Obesity and impaired metabolic health in patients with COVID-19. *Nat Rev Endocrinol*. 2020;16(7):341-342. doi:10.1038/s41574-020-0364-6

80. Zhu Z, Hasegawa K, Ma B, Fujiogi M, Camargo CA, Liang L. Association of obesity and its genetic predisposition with the risk of severe COVID-19: Analysis of population-based cohort data. *Metabolism*. 2020;112:154345. doi:10.1016/j.metabol.2020.154345

81. Aung N, Khanji MY, Munroe PB, Petersen SE. Causal Inference for Genetic Obesity, Cardiometabolic Profile and COVID-19 Susceptibility: A Mendelian Randomization Study. *Front Genet*. 2020;11. doi:10.3389/fgene.2020.586308

82. Kasela S, Ortega VE, Martorella M, et al. Genetic and non-genetic factors affecting the expression of COVID-19-relevant genes in the large airway epithelium. *Genome Med*. 2021;13(1):66. doi:10.1186/s13073-021-00866-2

83. Salvatore M, Gu T, Mack JA, et al. A phenome-wide association study (PheWAS) of COVID-19 outcomes by race using the electronic health records data in Michigan Medicine. *medRxiv*. Published online July 1, 2020:2020.06.29.20141564. doi:10.1101/2020.06.29.20141564

84. Oetjens MT, Luo JZ, Chang A, et al. Electronic health record analysis identifies kidney disease as the leading risk factor for hospitalization in confirmed COVID-19 patients. *PLoS One*. 2020;15(11 November). doi:10.1371/journal.pone.0242182

85. Joint United Nations Programme on HIV/AIDS (UNAIDS). Global AIDS Update 2018: Miles to go, closing gaps, breaking barriers, righting injustices. Published online 2018. http://www.unaids.org/en/resources/documents/2018/global-aids-update

86. Joint United Nations Programme on HIV/AIDS (UNAIDS). Country factsheets: Mozambique. Published 2017. http://www.unaids.org/en/regionscountries/countries/mozambique

87. Ministério da Saúde- MISAU, Instituto Nacional de Estatística - INE, ICF. *Inquérito de Indicadores de Imunização, Malária e HIV/SIDA Em Moçambique (IMASIDA) 2015*. MISAU/Moçambique, INE, and ICF; 2018. http://dhsprogram.com/pubs/pdf/AIS12/AIS12.pdf

88. Department of Defense HIV/AIDS Prevention Program. HIV Seroprevalence and Behavioral Epidemiology Risk Survey (SABERS): Mozambique 2016. Published online 2016.

89. Mozambique: 11 percent of military infected with HIV/AIDS. *APA News*. https://clubofmozambique.com/news/mozambique-11-percent-of-military-infected-with-hivaids/. Published 2018.

90. Djibo DA, Sahr F, McCutchan JA, et al. Prevalence and Risk Factors for Human Immunodeficiency Virus (HIV) and Syphilis Infections Among Military Personnel in Sierra Leone. *Curr HIV Res*. 2017;15(2). doi:10.2174/1570162x15666170517101349

91. Chretien JP, Blazes DL, Coldren RL, et al. The importance of militaries from developing countries in global infectious disease surveillance. *Bull World Health Organ*. 2007;85(3):174-180. doi:10.2471/BLT.06.037101

92. Endres-Dighe S, Farris T, Courtney L. Lessons learned from twelve years of HIV Seroprevalence and Behavioral Epidemiology Risk Survey (SABERS) development and implementation among foreign militaries. Riddle MS, ed. *PLoS One*. 2018;13(9):e0203718. doi:10.1371/journal.pone.0203718

93. Asefnia N, Cowan L, Werth R. HIV Risk Behavior and Prevention Considerations Among Military Personnel in Three Caribbean Region Countries: Belize, Barbados, and the Dominican Republic. *Curr HIV Res*. 2017;15(3). doi:10.2174/1570162x15666170517121316

94. Nacional De Prevalência I. *INSIDA 2009 Relatório Final*.; 2009. Accessed May 1, 2021. http://www.measuredhs.com

95. Asefnia N, Cowan L, Werth R. HIV risk behavior and prevention considerations among military personnel in three Caribbean region countries: Belize, Barbados, and the Dominican Republic. *Curr HIV Res*. 2017;15(3).

96. Djibo D, Sahr F, McCutchan A, Jain S, Araneta MRG. Prevalence and risk factors for human immunodeficiency virus (HIV) and syphilis infections among military personnel in Sierra Leone. *Curr HIV Res*. 2017;15.

97. Anastario MP, Tavarez MI, Chun H. Sexual risk behavior among military personnel stationed at border-crossing zones in the Dominican Republic. *Rev Panam Salud Publica/Pan Am J Public Heal*. 2010;28(5):361-367. doi:10.1590/S1020-49892010001100006

98. Bing EG, Ortiz DJ, Ovalle-Bahamón RE, et al. HIV/AIDS behavioral surveillance among Angolan military men. *AIDS Behav*. 2008;12(4):578-584. doi:10.1007/s10461-007-9280-1

99. Nwokoji U, Ajuwon A. Knowledge of AIDS and HIV risk-related sexual behavior among Nigerian naval personnel. *BMC Public Health*. 2004;4.

100. Abeler-Dörner L, Grabowski MK, Rambaut A, Pillay D, Fraser C. PANGEA-HIV 2: Phylogenetics and Networks for Generalised Epidemics in Africa. *Curr Opin HIV AIDS*. 2019;14(3):173-180. doi:10.1097/COH.0000000000000542

101. Kiwuwa-Muyingo S, Nazziwa J, Ssemwanga D, et al. HIV-1 transmission networks in high risk fishing communities on the shores of Lake Victoria in Uganda: A phylogenetic and epidemiological approach. *PLoS One*. 2017;12(10). doi:10.1371/journal.pone.0185818

102. Grabowski MK, Lessler J, Redd AD, et al. The Role of Viral Introductions in Sustaining Community-Based HIV Epidemics in Rural Uganda: Evidence from Spatial Clustering, Phylogenetics, and Egocentric Transmission Models. *PLoS Med*. 2014;11(3). doi:10.1371/journal.pmed.1001610

103. Okano JT, Sharp K, Valdano E, Palk L, Blower S. HIV transmission and source–sink dynamics in sub-Saharan Africa. *Lancet HIV*. 2020;7(3):e209-e214. doi:10.1016/S2352-3018(19)30407-2

104. Dellar RC, Dlamini S, Karim QA. Adolescent girls and young women: Key populations for HIV epidemic control. *J Int AIDS Soc*. 2015;18(2):64-70. doi:10.7448/IAS.18.2.19408

105. Karim SSA, Karim QA. *HIV/AIDS in South Africa*. 2nd ed. Cambridge University Press; 2010. doi:DOI: 10.1017/CBO9781139062404

106. Chapman J, Do Nascimento N, Mandal M. Role of male sex partners in HIV risk of adolescent girls and young women in Mozambique. *Glob Heal Sci Pract*. 2019;7(3):435-446. doi:10.9745/GHSP-D-19-00117

107. Campbell ANC, Tross S, Dworkin SL, et al. Relationship power and sexual risk among women in community-based substance abuse treatment. *J Urban Heal*. 2009;86(6):951-964. doi:10.1007/s11524-009-9405-0

108. Pulerwitz J, Mathur S, Woznica D. How empowered are girls/young women in their sexual relationships? Relationship power, HIV risk, and partner violence in Kenya. *PLoS One*. 2018;13(7). doi:10.1371/journal.pone.0199733

109. Leclerc-Madlala S. Age-disparate and intergenerational sex in southern Africa: The dynamics of hypervulnerability. *AIDS*. 2008;22(SUPPL. 4). doi:10.1097/01.aids.0000341774.86500.53

110. Higgins JA, Hoffman S, Dworkin SL. Rethinking gender, heterosexual men, and women's vulnerability to HIV/AIDS. *Am J Public Health*. 2010;100(3):435-445. doi:10.2105/AJPH.2009.159723

111. Mavedzenge SN, Weiss HA, Montgomery ET, et al. Determinants of differential HIV incidence among women in three southern African locations. *J Acquir Immune Defic Syndr*. 2011;58(1):89-99. doi:10.1097/QAI.0b013e3182254038

112. Kyegombe N, Meiksin R, Wamoyi J, Heise L, Stoebenau K, Buller AM. Sexual health of adolescent girls and young women in Central Uganda: exploring perceived coercive aspects of transactional sex. *Sex Reprod Heal Matters*. 2020;28(1). doi:10.1080/26410397.2019.1700770

113. Agency CI. *The World Factbook 2016-2017*.; 2016. https://www.cia.gov/library/publications/the-world-factbook/index.html

114. Research Electronic Data Capture (REDCap). Published 2018. https://www.project-redcap.org/

115. Esri Inc. ArcGIS Pro (Version 2.7). Published online 2021. https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview

116. Esri - Garmin International Inc. World Water Bodies. Published 2018. https://www.arcgis.com/home/item.html?id=e750071279bf450cbd510454a80f2e63

117. GADM database of Global Administrative Areas. https://gadm.org/

118. contributors O. No Title. Published 2015. http://download.geofabrik.de/africa/mozambique.html

119. Makanga PT, Schuurman N, Sacoor C, et al. Seasonal variation in geographical access to maternal health services in regions of southern Mozambique. *Int J Health Geogr*. 2017;16(1):1. doi:10.1186/s12942-016-0074-4

120. Babor T, Higgins-Biddle J, Saunders J, Monteiro M. *AUDIT: The Alcohol Use Disorders Identification Test, Guidelines for Use in Primary Care (Second Edition)*. Second Edi. World Health Organization; 2001. https://apps.who.int/iris/bitstream/handle/10665/67205/WHO_MSD_MSB_01.6a.pdf;jsessionid=6ADBC000990EDC5BB4BE9380DEEDCD94?sequence=1

121. Moon TD, Ossemane EB, Green AF, et al. Antiretroviral therapy program expansion in Zambézia Province, Mozambique: geospatial mapping of community-based and health facility data for integrated health planning. *PLoS One*. 2014;9(10):e109653-e109653. doi:10.1371/journal.pone.0109653

122. Dos Anjos Luis A, Cabral P. Geographic accessibility to primary healthcare centers in Mozambique. *Int J Equity Health*. 2016;15(1):173. doi:10.1186/s12939-016-0455-0

123. Cuadros DF, Li J, Branscum AJ, et al. Mapping the spatial variability of HIV infection in Sub-Saharan Africa: Effective information for localized HIV prevention and control. *Sci Rep*. 2017;7(1):9093. doi:10.1038/s41598-017-09464-y

Appendices

Supplementary Table 1: Odds Ratio (OR) and 95% Confidence Interval (95% CI) for severe COVID-19 infection by comorbidity Latent Class membership among participants over the age of 65 years, adjusting for participant age and sex.

Latent Class	Researcher Label	# of Participants *N (%)*	COVID-19 Outcome *N (%)*		OR (95% CI)[a]	
			Hospital[b]	Mortality[c]	Hospital[b]	Mortality[c]
1	*"Healthy"*	125,564 (73.5)	287 (0.2)	72 (0.1)	Reference	Reference
2	*"Some Non-Specific Health Conditions"*	29,006 (17.0)	202 (0.7)	118 (0.4)	3.5 (2.6-4.7)	12.4 (5.0-30.6)
3	*"Diabetics with 1-2 other conditions"*	10,116 (5.9)	96 (0.9)	57 (0.6)	5.1 (3.7-7.0)	17.6 (7.4-41.5)
4	*"Cardiac multimorbidity"*	4,033 (2.4)	91 (2.3)	56 (1.4)	12.9 (9.7-17.2)	44.8 (19.6-102.0)
5	*"Cancer multimorbidity"*	2,015 (1.2)	27 (1.3)	13 (0.6)	8.3 (5.4-12.7)	24.5 (9.2-64.9)

[a] ORs & 95% CIs defined from multinomial logistic regression model with the outcome coded as 'No Event' (Reference), 'COVID-19 Hospitalization without Mortality', and 'COVID-19 Mortality (with or without hospitalization)', adjusting for age and sex.
[b] Hospital inpatient diagnosis (primary or secondary) of ICD10 code U07.1 (lab-confirmed COVID-19) or U07.2 (clinically/epidemiologically-diagnosed COVID-19) without mortality event after January 1st, 2020 with primary or contributing cause recorded with ICD-10 U07.1 or U07.2 codes.
[c] Mortality event after January 1st, 2020 with primary or contributing cause recorded with ICD-10 U07.1 or U07.2 codes

Supplementary Table 2: Odds Ratio (OR) and 95% Confidence Interval (95% CI) for severe COVID-19 infection by comorbidity Latent Class membership stratified by pandemic phase, adjusting for participant age and sex.

Latent Class	Researcher Label	# of Participants *N (%)*	Severe COVID-19 Infection *N (%)*		OR (95% CI)[a]	
			Phase 1[b]	Phase 2[c]	Phase 1[b]	Phase 2[c]
1	*"Healthy"*	72,325 (67.2)	165 (0.1)	194 (0.2)	Reference	Reference
2	*"Some Non-Specific Health Conditions"*	22,985 (21.3)	159 (0.5)	161 (0.6)	4.9 (3.3- 7.2)	4.0 (2.8- 5.7)
3	*"Diabetics with 1-2 other conditions"*	7,719 (7.2)	60 (0.6)	93 (0.9)	5.0 (3.2- 7.8)	7.5 (5.2-10.6)
4	*"Cardiac multimorbidity"*	3,348 (3.1)	73 (1.8)	74 (1.8)	18.4 (12.8-26.5)	15.0 (10.7-21.1)
5	*"Cancer multimorbidity"*	1,325 (1.2)	17 (0.8)	23 (1.1)	10.7 (6.2-18.2)	9.2 (5.5-15.5)

[a] ORs & 95% CIs defined from multinomial logistic regression model with the outcome coded as 'No Event' (Reference), 'Phase 1', and 'Phase 2', adjusting for age and sex.
[a] Phase 1 defined as events prior to 01-July-2020
[b] Phase 2 defined as events on or after 01-July-2020

Supplementary Table 3: Odds Ratio (OR) and 95% Confidence Interval (95% CI) for severe COVID-19 infection by comorbidity Latent Class membership among participants 65 years or younger, adjusting for participant age and sex.

Latent Class	**Researcher Label**	**# of Participants** *N (%)*	**Severe COVID-19 Infection** *N (%)*	**OR (95% CI)**
1	*"Healthy"*	53,239 (84.5)	104 (0.2)	Reference
2	*"Some Non-Specific Health Conditions"*	6,021 (9.6)	36 (0.6)	2.8 (1.3- 5.7)
3	*"Diabetics with 1-2 other conditions"*	2,397 (3.8)	29 (1.2)	7.6 (4.4-12.9)
4	*"Cardiac multimorbidity"*	685 (1.1)	19 (2.8)	21.1 (12.9-34.4)
5	*"Cancer multimorbidity"*	690 (1.1)	7 (1.0)	9.6 (4.2-21.6)

Supplementary Table 4: Odds Ratio (OR) and 95% Confidence Interval (95% CI) for severe COVID-19 infection by comorbidity Latent Class membership among participants over the age of 65 years, adjusting for participant age and sex.

Latent Class	**Researcher Label**	**# of Participants** *N (%)*	**Severe COVID-19 Infection** *N (%)*	**OR (95% CI)**
1	*"Healthy"*	72,325 (67.2)	255 (0.4)	Reference
2	*"Some Non-Specific Health Conditions"*	22,985 (21.3)	284 (1.2)	4.6 (3.4- 6.4)
3	*"Diabetics with 1-2 other conditions"*	7,719 (7.2)	124 (1.6)	6.1 (4.4- 8.5)
4	*"Cardiac multimorbidity"*	3,348 (3.1)	128 (3.8)	15.9 (11.9-21.5)
5	*"Cancer multimorbidity"*	1,325 (1.2)	33 (2.5)	10.0 (6.5-15.4)

Supplementary Table 5: Odds Ratio (OR) and 95% Confidence Interval (95% CI) for severe COVID-19 infection by comorbidity Latent Class membership in Primary, Unrelated, and Mixed Kinship samples, adjusting for participant age and sex.

Latent Class	**Researcher Label**	**Primary Sample** *(n=170,734)* *OR (95% CI)*	**'Unrelated' Sample** (n=151,623) *OR (95% CI)*	**'Mixed Kinship' Sample** (n=151,623) *OR (95% CI)*
1	*"Healthy"*	Reference	Reference	Reference
2	*"Some Non-Specific Health Conditions"*	4.39 (3.36, 5.74)	4.57 (3.44, 6.05)	4.57 (3.44, 6.08)
3	*"Diabetics with 1-2 other conditions"*	6.36 (4.82, 8.40)	6.37 (4.75, 8.54)	6.75 (5.03, 9.05)
4	*"Cardiac multimorbidity"*	16.55 (12.89,21.23)	16.31 (12.45,21.37)	17.18 (13.13,22.48)
5	*"Cancer multimorbidity"*	9.88 (6.80,14.34)	10.23 (6.93,15.10)	11.12 (7.56,16.36)

Supplementary Table 6: Odds Ratio (OR) and 95% Confidence Intervals (95% CI) for severe COVID-19 infection by estimation method.

	Simultaneous Estimate[a]		**Disease-by-Disease[b]**	
Elixhauser Comorbidity	Without '# of Comorbidities'[c] *OR (95% CI)*	With '# of Comorbidities'[d] *OR (95% CI)*	Without '# of Comorbidities'[c] *OR (95% CI)*	With '# of Comorbidities'[d] *OR (95% CI)*
Alcohol abuse	**1.48 (1.10,2.00)****	**1.39 (1.04,1.85)***	**2.85 (2.14, 3.72)*****	1.55 (1.16,2.02)
Anemia, deficiency	1.16 (0.90,1.49)	1.14 (0.89,1.46)	**2.77 (2.17, 3.48)*****	**1.48 (1.16,1.87)***
Anemia, blood loss	-		-	-
Cardiac arrhythmia	**1.24 (1.04,1.48)***	1.12 (0.94,1.33)	**2.29 (1.96, 2.67)*****	1.19 (1.01,1.39)
Congestive heart failure	0.94 (0.72,1.24)	1.00 (0.77,1.31)	**2.92 (2.30, 3.64)*****	**1.54 (1.21,1.93)****
Coagulopathy	1.00 (0.62,1.62)	1.01 (0.63,1.61)	**2.76 (1.70, 4.20)*****	1.51 (0.93,2.31)
Depression	**1.52 (1.22,1.89)*****	**1.41 (1.14,1.74)****	**3.19 (2.60, 3.87)*****	**1.63 (1.32,1.99)*****
Diabetes, complicated	1.22 (0.91,1.64)	1.23 (0.92,1.65)	**3.28 (2.50, 4.23)*****	**1.66 (1.26,2.15)****
Diabetes, uncomplicated	**1.70 (1.45,2.00)*****	**1.48 (1.26,1.75)*****	**2.64 (2.29, 3.04)*****	**1.43 (1.23,1.65)*****
Drug abuse	1.15 (0.35,3.71)	1.11 (0.35,3.56)	3.49 (0.86, 9.26)	1.59 (0.39,4.20)
Fluid/electrolyte disorders	**1.71 (1.38,2.12)*****	**1.74 (1.41,2.15)*****	**4.96 (4.11, 5.95)*****	**2.66 (2.19,3.20)*****
HIV/AIDS	-		-	-
Hypertension, complicated	**2.38 (1.05,5.39)***	**2.34 (1.04,5.24)***	**7.08 (2.97,14.24)*****	**3.72 (1.56,7.46)***
Hypertension, uncomplicated	**1.55 (1.34,1.79)*****	**1.21 (1.03,1.42)***	**2.70 (2.38, 3.07)*****	1.17 (1.01,1.36)
Hypothyroidism	1.19 (0.94,1.51)	1.08 (0.85,1.37)	**2.06 (1.62, 2.58)*****	1.07 (0.84,1.34)
Liver disease	**1.70 (1.28,2.27)*****	**1.66 (1.26,2.20)*****	**4.14 (3.16, 5.33)*****	**2.13 (1.62,2.75)*****
Lymphoma	**1.60 (1.01,2.53)***	1.45 (0.92,2.28)	**2.74 (1.71, 4.13)*****	1.55 (0.97,2.34)
Metastatic cancer	**1.96 (1.37,2.80)*****	**1.82 (1.28,2.59)*****	**3.42 (2.47, 4.60)*****	**1.74 (1.25,2.34)***
Other neurological disorders	**2.70 (2.25,3.24)*****	**2.39 (1.99,2.87)*****	**4.24 (3.56, 5.01)*****	**2.53 (2.12,3.00)*****
Obesity	0.97 (0.81,1.17)	0.88 (0.74,1.06)	**1.63 (1.37, 1.93)*****	0.84 (0.70,1.00)
Paralysis	**2.09 (1.45,3.01)*****	**2.07 (1.44,2.97)*****	**6.05 (4.22, 8.40)*****	**3.24 (2.25,4.50)*****
Peptic ulcer disease	0.91 (0.57,1.46)	0.87 (0.54,1.38)	1.69 (1.03, 2.59)	0.94 (0.57,1.44)
Peripheral vascular disease	1.08 (0.82,1.41)	1.07 (0.82,1.40)	**2.59 (1.99, 3.30)*****	1.40 (1.07,1.79)
Psychoses	1.44 (0.80,2.57)	1.33 (0.75,2.35)	**3.31 (1.80, 5.51)*****	1.75 (0.95,2.93)
Chronic pulmonary disease	**1.59 (1.38,1.84)*****	**1.37 (1.18,1.59)*****	**2.21 (1.92, 2.53)*****	1.26 (1.09,1.45)
Pulmonary circulation disorder	1.07 (0.70,1.64)	1.05 (0.69,1.61)	**2.57 (1.66, 3.79)*****	1.38 (0.89,2.03)
Rheumatoid arthritis	**1.48 (1.15,1.90)****	**1.36 (1.06,1.75)***	**2.42 (1.89, 3.06)*****	1.33 (1.03,1.69)
Renal failure	**1.50 (1.21,1.87)*****	**1.51 (1.22,1.87)*****	**3.72 (3.05, 4.49)*****	**1.98 (1.62,2.41)*****
Solid tumor without metastasis	**1.29 (1.05,1.58)***	1.14 (0.93,1.40)	**2.04 (1.70, 2.43)*****	1.14 (0.95,1.36)
Valvular disease	1.14 (0.88,1.48)	1.11 (0.86,1.44)	**2.54 (2.00, 3.19)*****	1.35 (1.05,1.69)
Weight loss	1.12 (0.79,1.58)	1.06 (0.75,1.49)	**2.49 (1.77, 3.41)*****	1.37 (0.97,1.87)

* p-value < 0.05; ** p-value < 0.01; *** p-value < 0.001

[a] ORs & 95% CIs for '*severe COVID-19 infection*' defined from logistic regression model with 29 of the 31 Elixhauser comorbidities, age, and sex as covariates

[b] ORs and 95% CIs for '*severe COVID-19 infection*' from n=29 separate logistic regression models, with each comorbidity as the exposure, adjusting for age and sex as covariates. Bonferroni correction applied to p-values for n=29 multiple tests.

[c] Only includes age and sex as covariates.

[d] Includes '# of Comorbidities' ('0','1','2 or more') as an additional categorical covariate with age and sex.

Supplementary Table 7: Odds Ratio (OR) and 95% Confidence Intervals (95% CI) for COVID-19 mortality by estimation method.

	Simultaneous Estimate[a]		**Disease-by-Disease[b]**	
Elixhauser Comorbidity	Without '# of Comorbidities'[c] *OR (95% CI)*	With '# of Comorbidities'[d] *OR (95% CI)*	Without '# of Comorbidities'[c] *OR (95% CI)*	With '# of Comorbidities'[d] *OR (95% CI)*
Alcohol abuse	**2.00 (1.28, 3.15)****	**1.81 (1.17, 2.81)****	**4.10 (2.65, 6.06)*****	**2.14 (1.38, 3.18)***
Anemia, deficiency	1.15 (0.76, 1.74)	1.13 (0.75, 1.68)	**2.98 (2.00, 4.28)*****	1.50 (1.01, 2.17)
Anemia, blood loss	-	-	-	-
Cardiac arrhythmia	1.22 (0.91, 1.63)	1.02 (0.77, 1.36)	**2.41 (1.86, 3.10)*****	1.14 (0.87, 1.47)
Congestive heart failure	1.01 (0.66, 1.54)	1.10 (0.72, 1.66)	**3.20 (2.21, 4.50)*****	1.58 (1.09, 2.23)
Coagulopathy	1.17 (0.58, 2.36)	1.21 (0.61, 2.38)	**3.85 (1.90, 6.90)****	2.04 (1.01, 3.65)
Depression	**1.67 (1.16, 2.40)****	**1.48 (1.04, 2.12)***	**3.93 (2.78, 5.41)*****	**1.92 (1.35, 2.66)****
Diabetes, complicated	1.26 (0.78, 2.04)	1.29 (0.80, 2.07)	**3.66 (2.34, 5.45)*****	1.76 (1.12, 2.64)
Diabetes, uncomplicated	**1.74 (1.32, 2.30)*****	**1.40 (1.06, 1.85)***	**2.81 (2.20, 3.57)*****	1.41 (1.10, 1.80)
Drug abuse	3.15 (0.72,13.76)	3.02 (0.71,12.81)	**10.18 (1.66,32.82)***	4.57 (0.75,14.68)
Fluid/electrolyte disorders	**1.51 (1.06, 2.15)***	**1.58 (1.13, 2.21)****	**5.30 (3.90, 7.07)*****	**2.65 (1.94, 3.55)*****
HIV/AIDS	-	-	-	-
Hypertension, complicated	2.40 (0.70, 8.19)	2.39 (0.72, 7.99)	**7.78 (1.90,21.00)***	3.87 (0.94,10.42)
Hypertension, uncomplicated	**1.61 (1.24, 2.08)*****	1.04 (0.80, 1.35)	**2.96 (2.36, 3.73)*****	1.06 (0.82, 1.37)
Hypothyroidism	0.86 (0.54, 1.35)	0.74 (0.47, 1.16)	1.61 (1.00, 2.43)	0.78 (0.49, 1.19)
Liver disease	**2.30 (1.48, 3.56)*****	**2.19 (1.43, 3.35)*****	**5.89 (3.90, 8.56)*****	**2.88 (1.90, 4.21)*****
Lymphoma	1.95 (0.99, 3.85)	1.68 (0.87, 3.27)	**3.66 (1.81, 6.54)****	1.97 (0.97, 3.53)
Metastatic cancer	1.81 (0.97, 3.39)	1.56 (0.84, 2.91)	**2.91 (1.58, 4.89)****	1.40 (0.76, 2.35)
Other neurological disorders	**3.89 (2.94, 5.13)*****	**3.13 (2.38, 4.13)*****	**6.22 (4.78, 8.00)*****	**3.55 (2.72, 4.59)*****
Obesity	0.85 (0.61, 1.19)	0.74 (0.53, 1.02)	1.47 (1.06, 1.99)	0.71 (0.51, 0.97)
Paralysis	**2.27 (1.32, 3.92)****	**2.22 (1.30, 3.80)****	**7.44 (4.28,12.02)*****	**3.79 (2.18, 6.14)*****
Peptic ulcer disease	0.62 (0.25, 1.55)	0.58 (0.24, 1.43)	1.26 (0.45, 2.74)	0.66 (0.24, 1.44)
Peripheral vascular disease	0.96 (0.61, 1.49)	0.96 (0.62, 1.48)	**2.44 (1.58, 3.62)*****	1.24 (0.80, 1.85)
Psychoses	1.96 (0.83, 4.67)	1.74 (0.74, 4.06)	**5.16 (2.03,10.68)****	2.65 (1.04, 5.50)
Chronic pulmonary disease	1.23 (0.95, 1.61)	1.01 (0.77, 1.31)	**1.86 (1.44, 2.38)*****	0.98 (0.75, 1.26)
Pulmonary circulation disorder	1.04 (0.50, 2.16)	0.98 (0.48, 2.03)	2.36 (1.07, 4.47)	1.19 (0.54, 2.26)
Rheumatoid arthritis	1.51 (0.99, 2.29)	1.32 (0.87, 1.99)	**2.49 (1.62, 3.66)*****	1.29 (0.84, 1.91)
Renal failure	**1.89 (1.35, 2.64)*****	**1.84 (1.34, 2.54)*****	**4.63 (3.42, 6.15)*****	**2.30 (1.70, 3.08)*****
Solid tumor without metastasis	1.21 (0.86, 1.71)	1.01 (0.72, 1.42)	**1.92 (1.40, 2.57)*****	1.00 (0.73, 1.35)
Valvular disease	1.04 (0.68, 1.59)	1.00 (0.66, 1.51)	**2.47 (1.65, 3.56)*****	1.22 (0.82, 1.77)
Weight loss	1.17 (0.67, 2.02)	1.07 (0.62, 1.83)	**2.77 (1.57, 4.51)****	1.44 (0.81, 2.34)

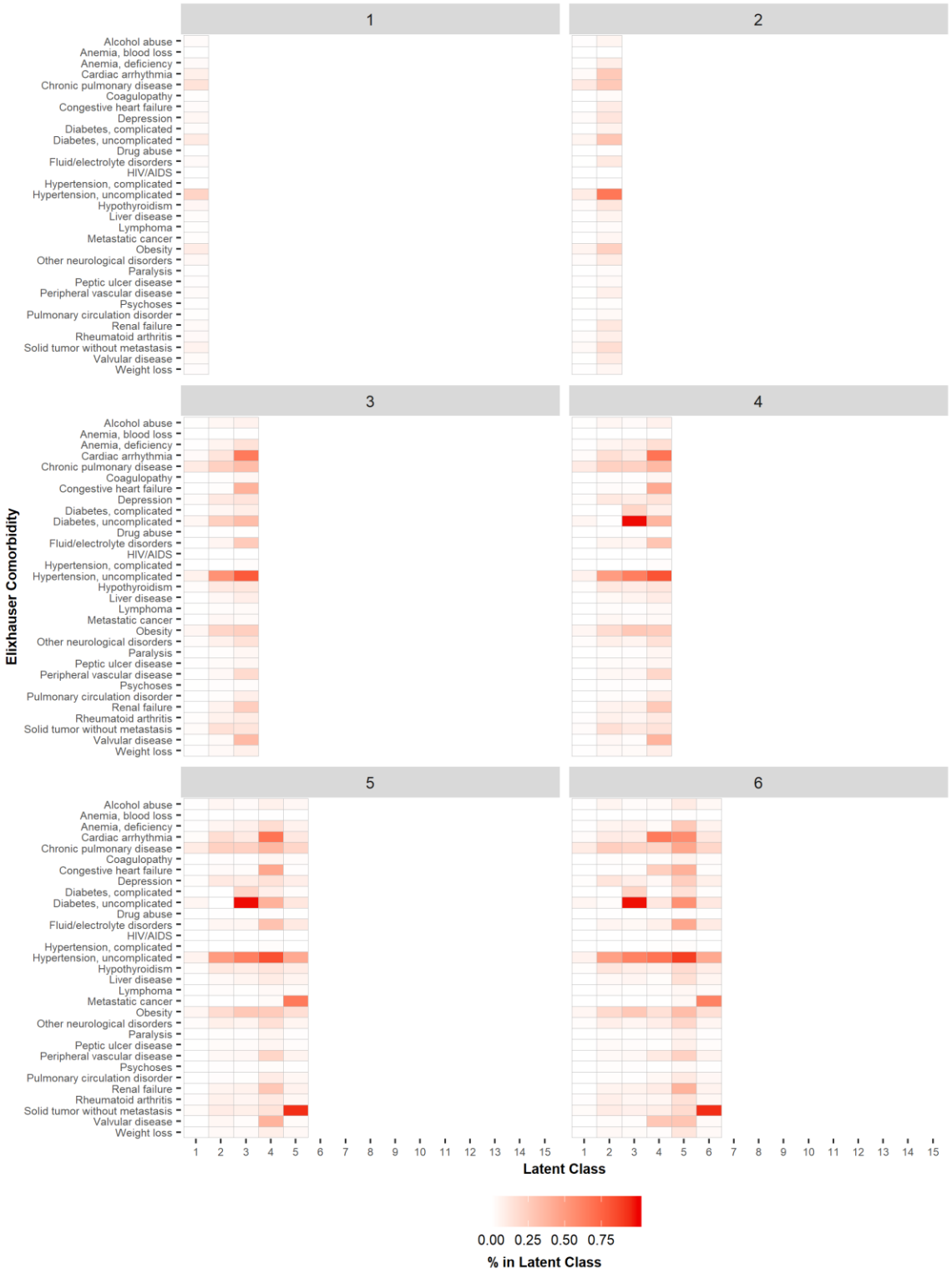
* p-value < 0.05; ** p-value < 0.01; *** p-value < 0.001

[a] ORs & 95% CIs for '*severe COVID-19 infection*' defined from logistic regression model with 29 of the 31 Elixhauser comorbidities, age, and sex as covariates

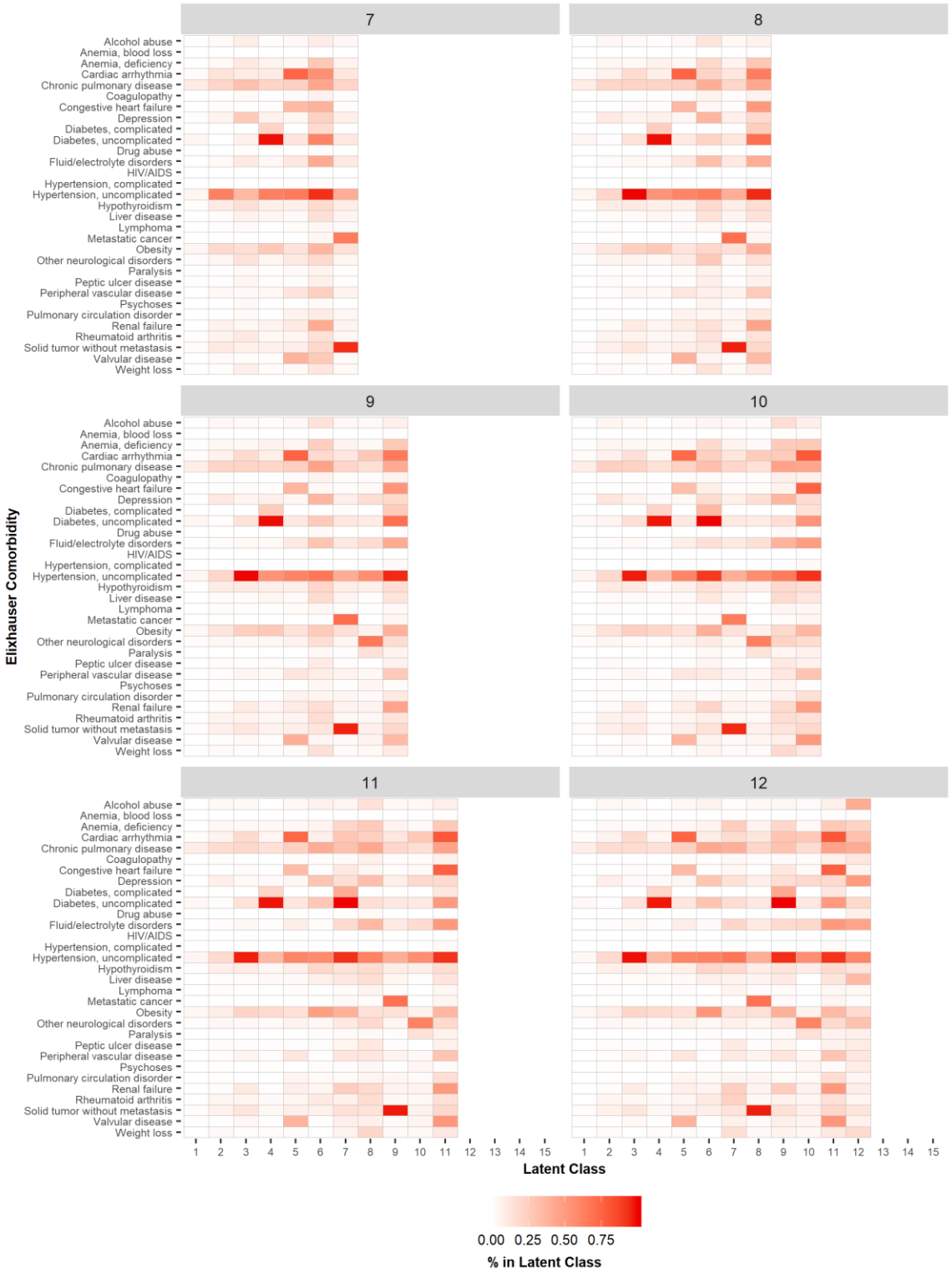
[b] ORs and 95% CIs for '*severe COVID-19 infection*' from n=29 separate logistic regression models, with each comorbidity as the exposure, adjusting for age and sex as covariates. Bonferroni correction applied to p-values for n=29 multiple tests.

[c] Only includes age and sex as covariates.

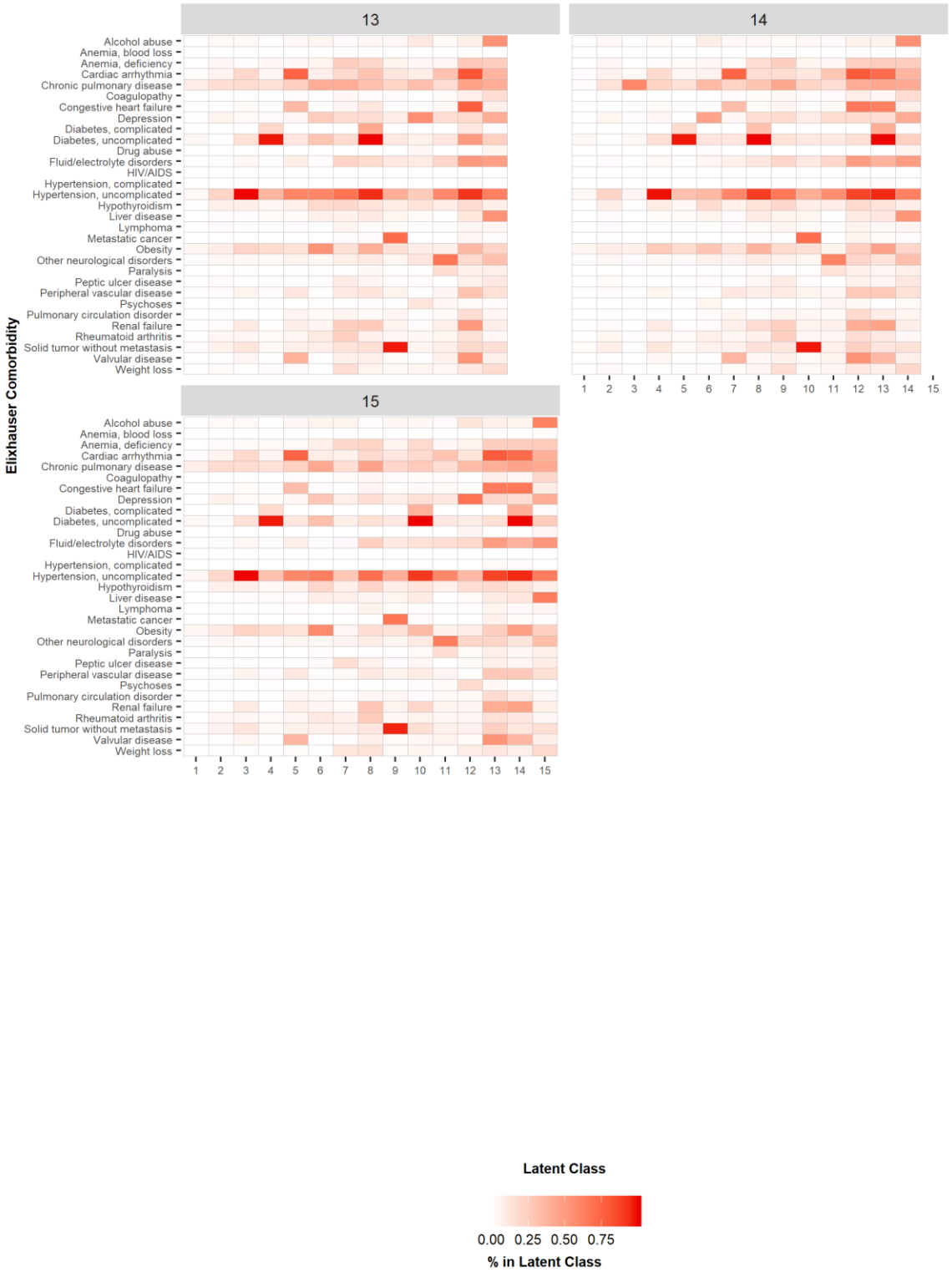
[d] Includes '# of Comorbidities' ('0','1','2 or more') as an additional categorical covariate with age and sex.



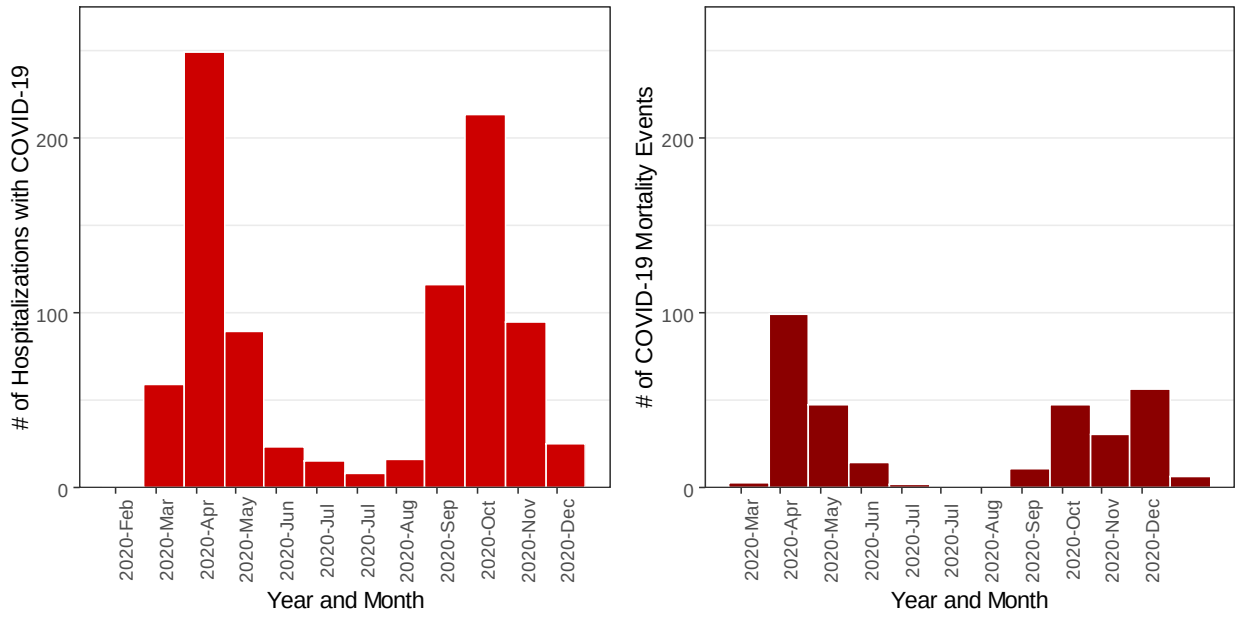
Supplementary Figure 1: Class-Specific Item Response Probabilities for 2-15 Class Solutions



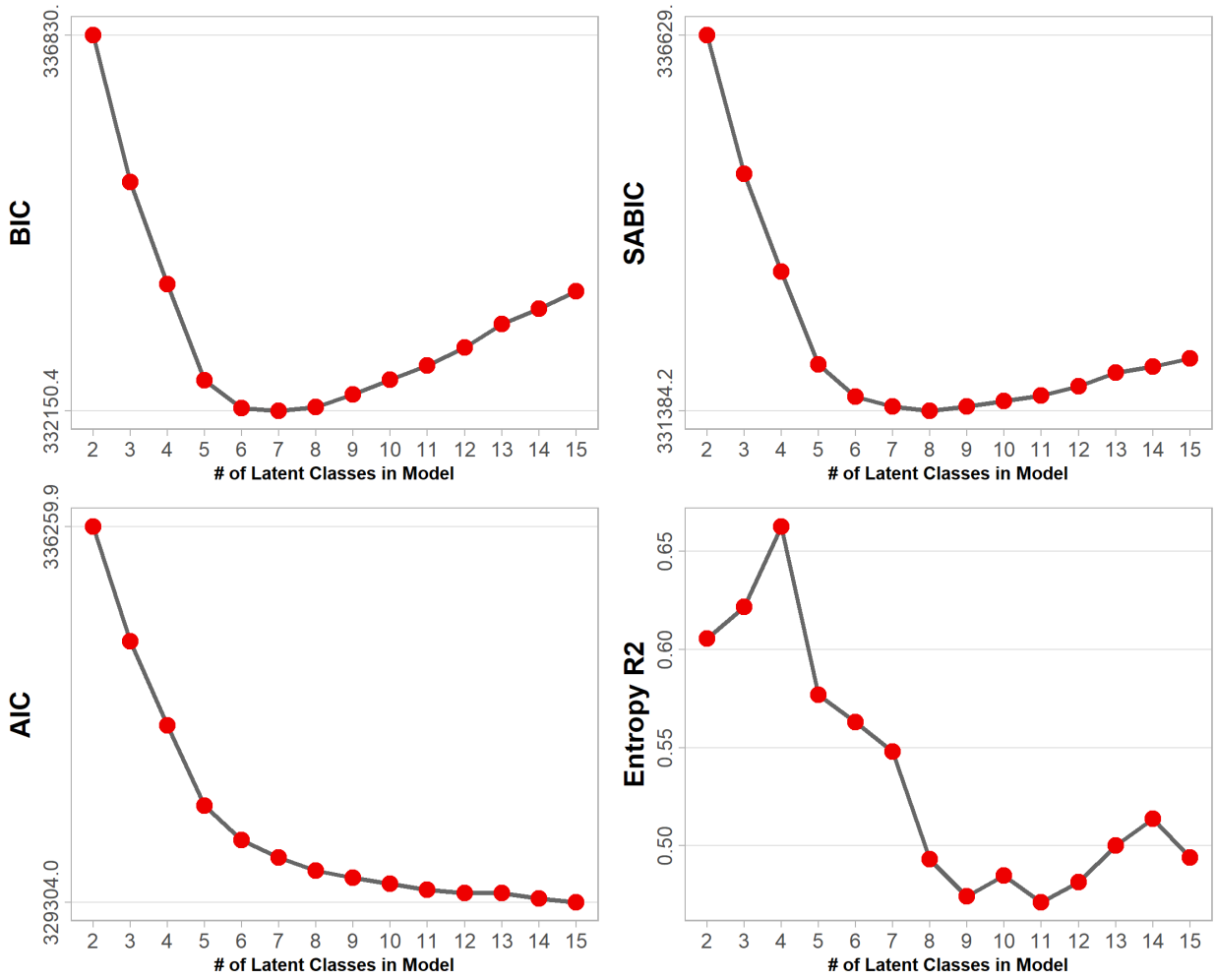
Supplementary Figure 1 (Continued): Class-Specific Item Response Probabilities for 2-15 Class Solutions



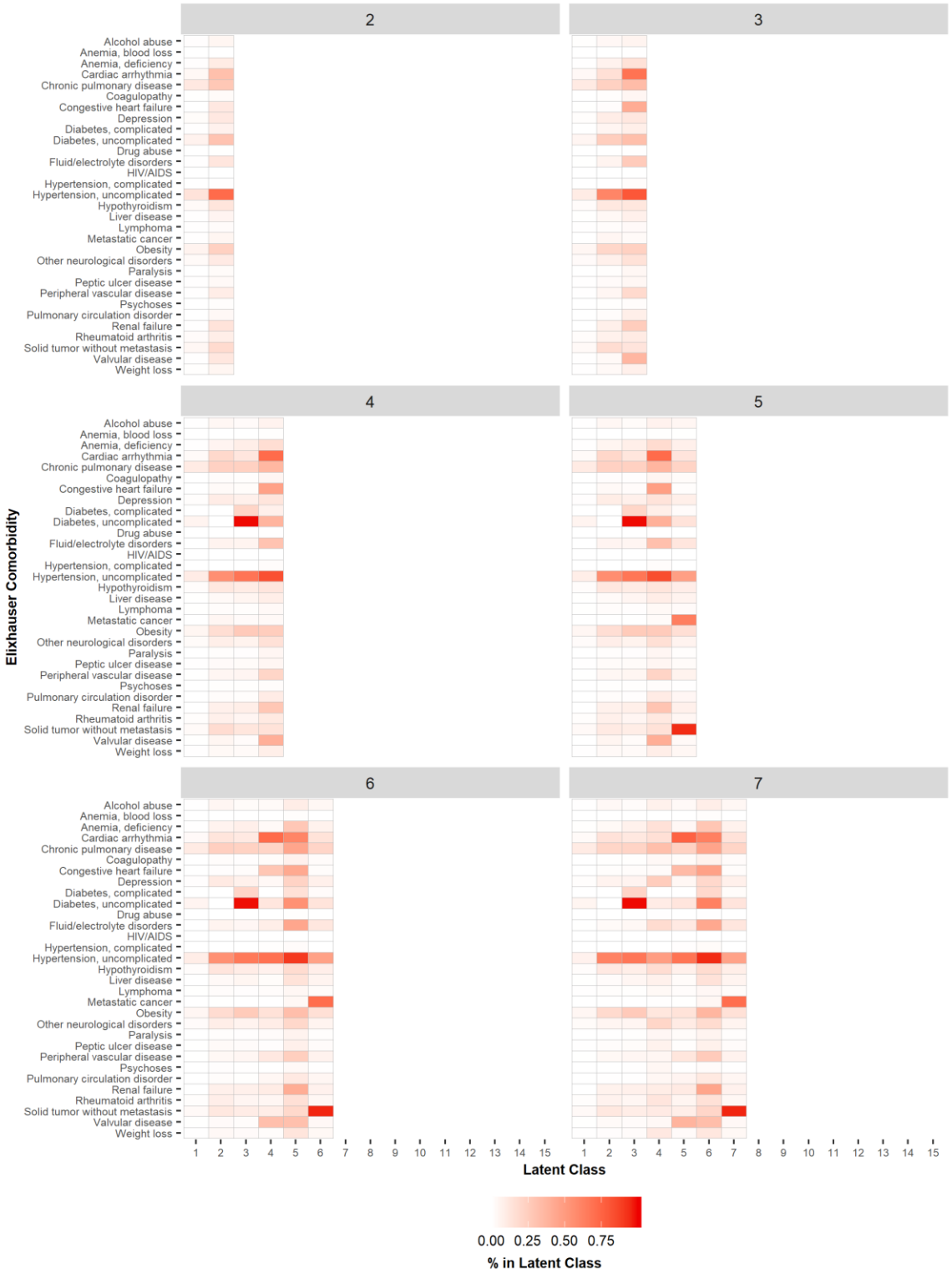
Supplementary Figure 1 (Continued): Class-Specific Item Response Probabilities for 2-15 Class Solutions



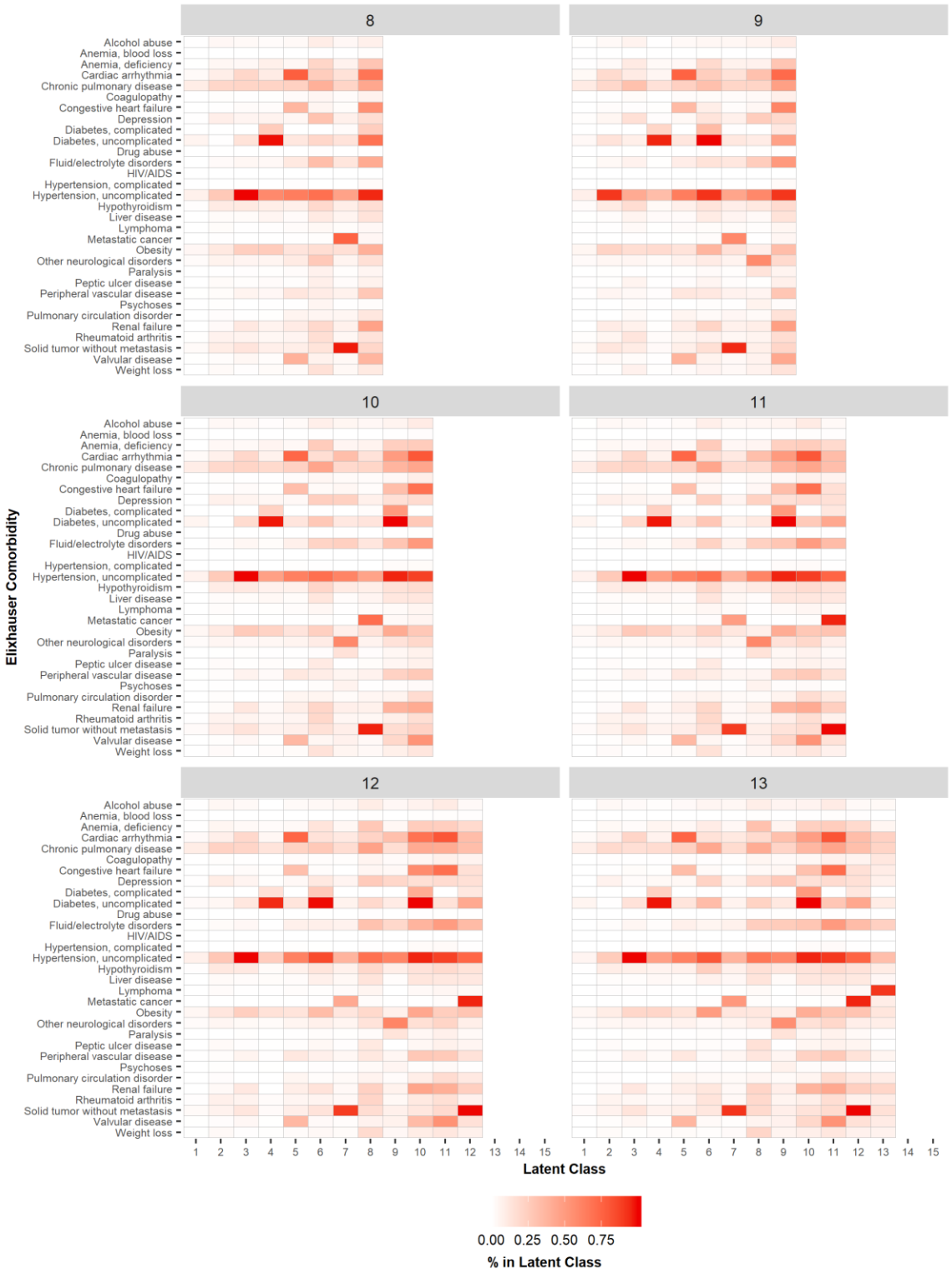
Supplementary Figure 2: Distribution of COVID-19 hospitalizations and mortality by Month and Year



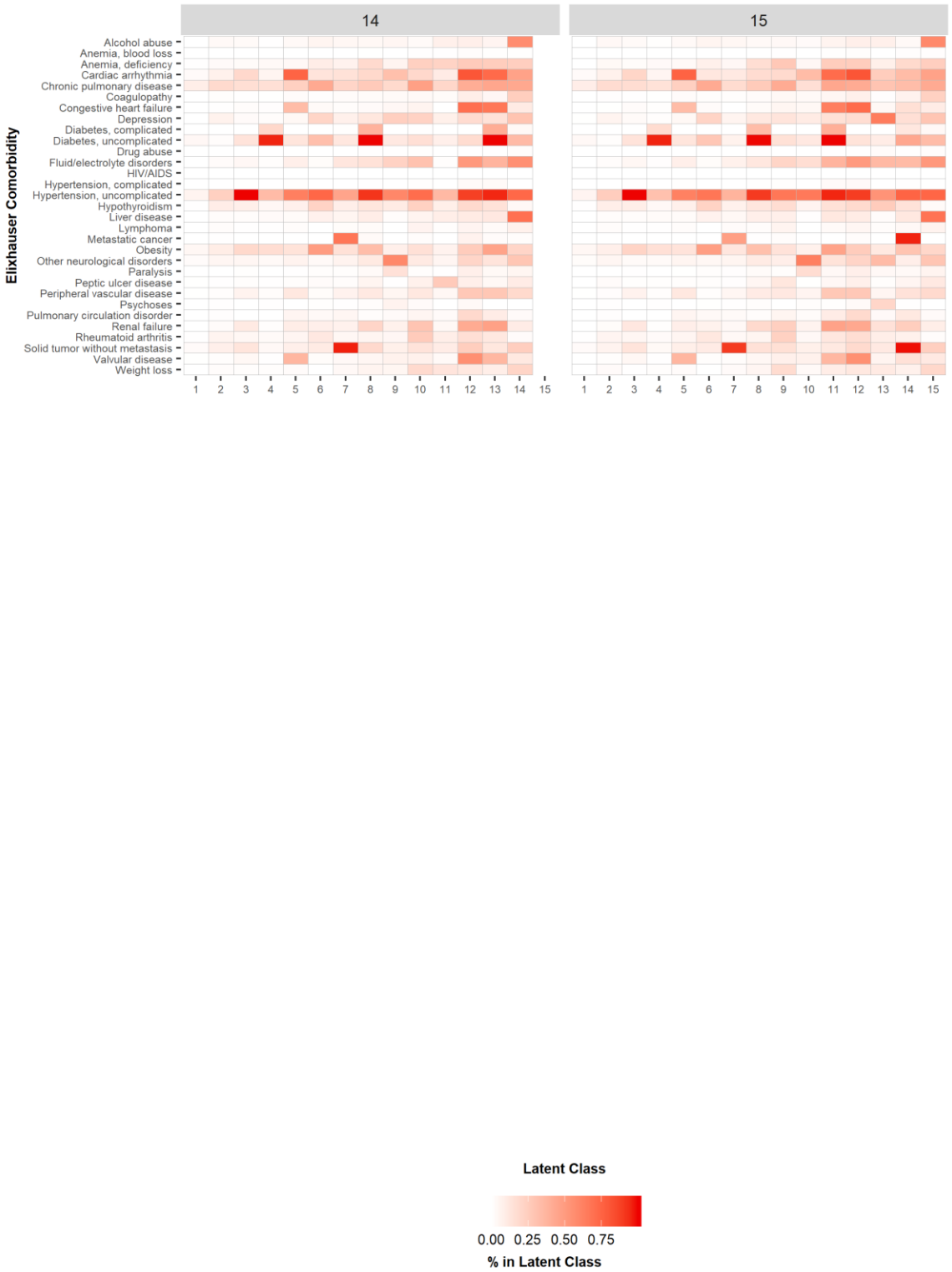
Supplementary Figure 3: Latent Class Analysis Model Results for 2-15 Class Solutions, '*Participants 65 Years or Younger*' sample (n=63,032)



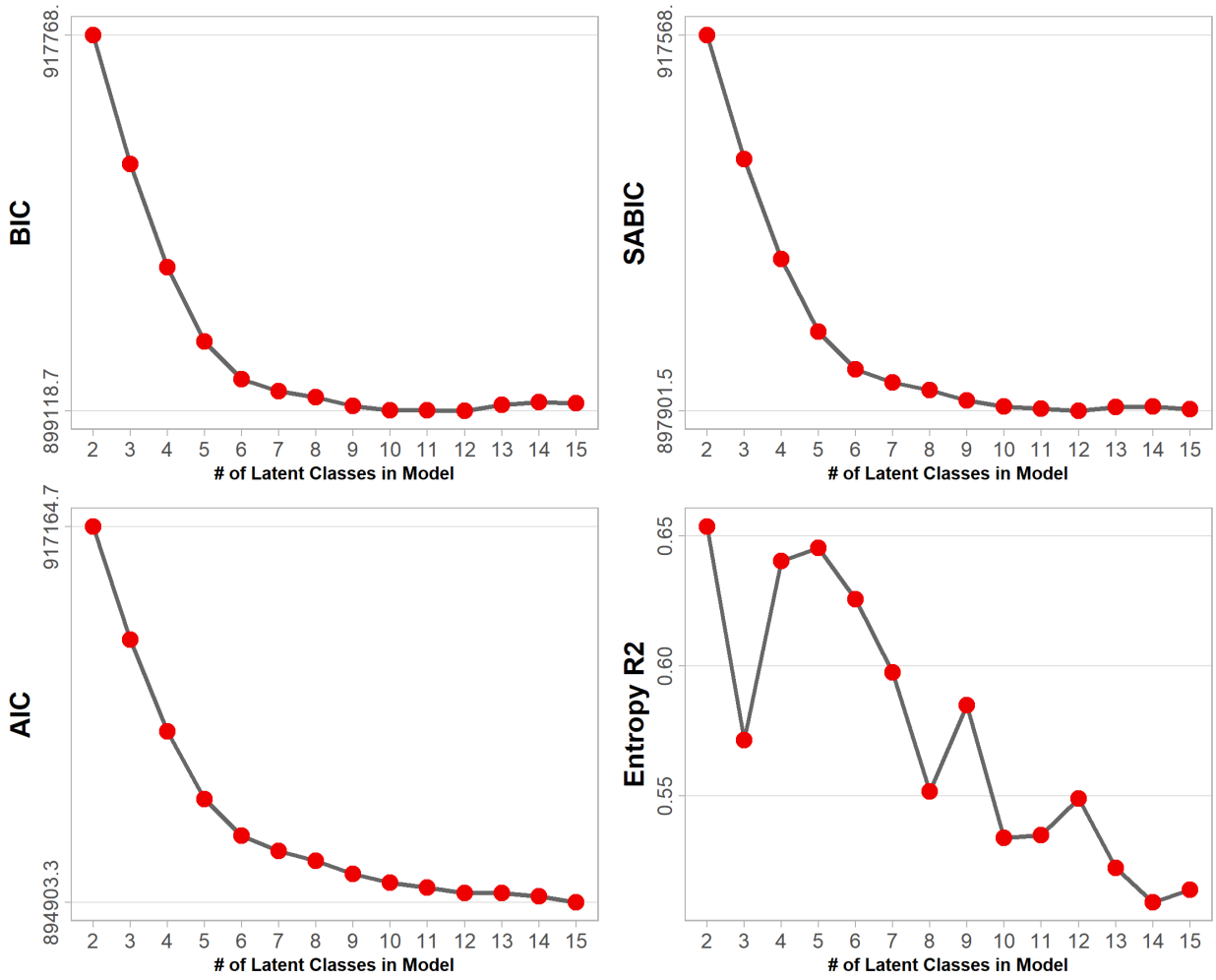
Supplementary Figure 4: Class-Specific Item Response Probabilities for 2-15 Class Solutions, '*Participants 65 Years or Younger*' sample (n=63,032)



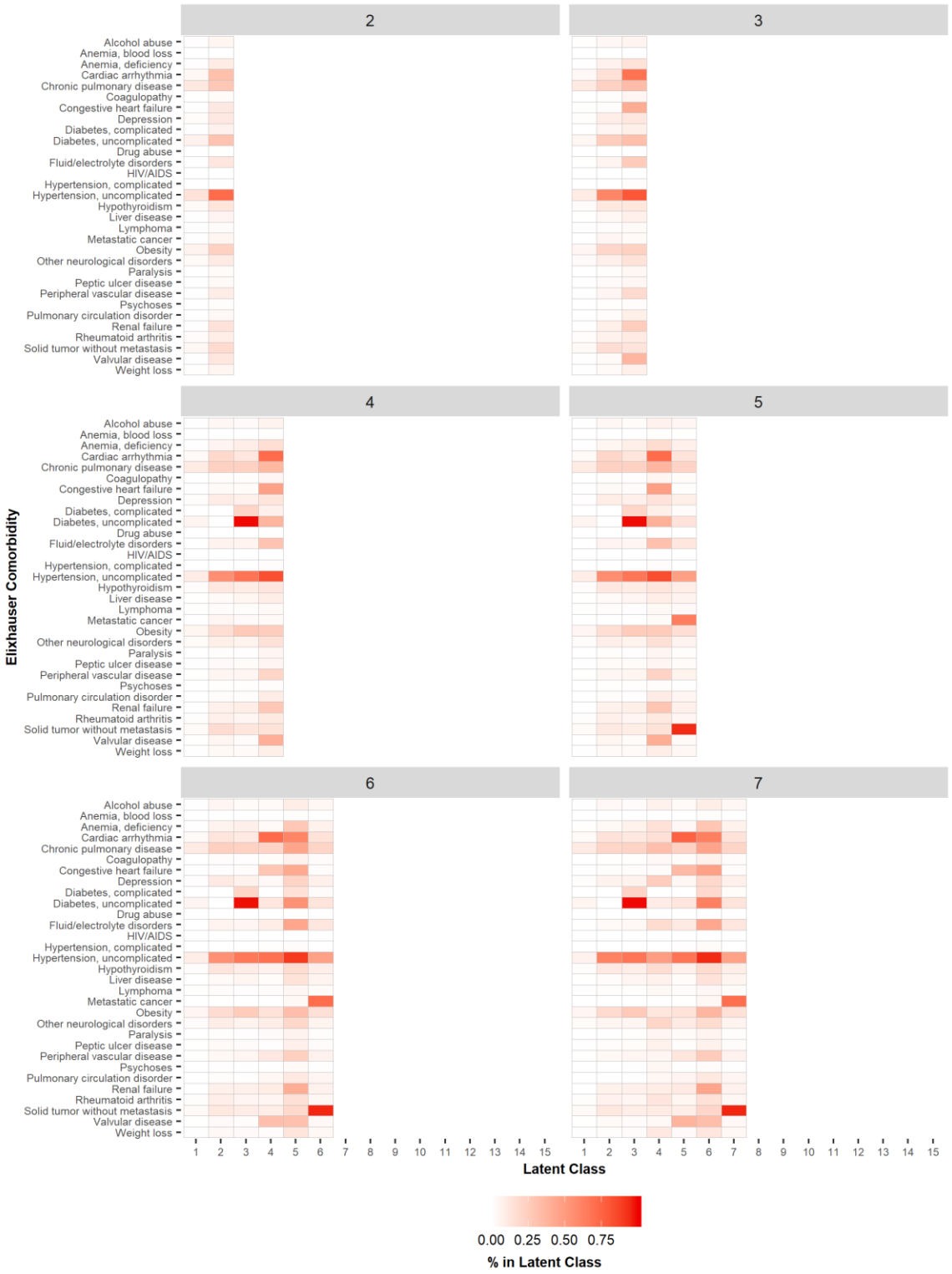
Supplementary Figure 4 (Continued): Class-Specific Item Response Probabilities for 2-15 Class Solutions, '*Participants 65 Years or Younger*' sample (n=63,032)



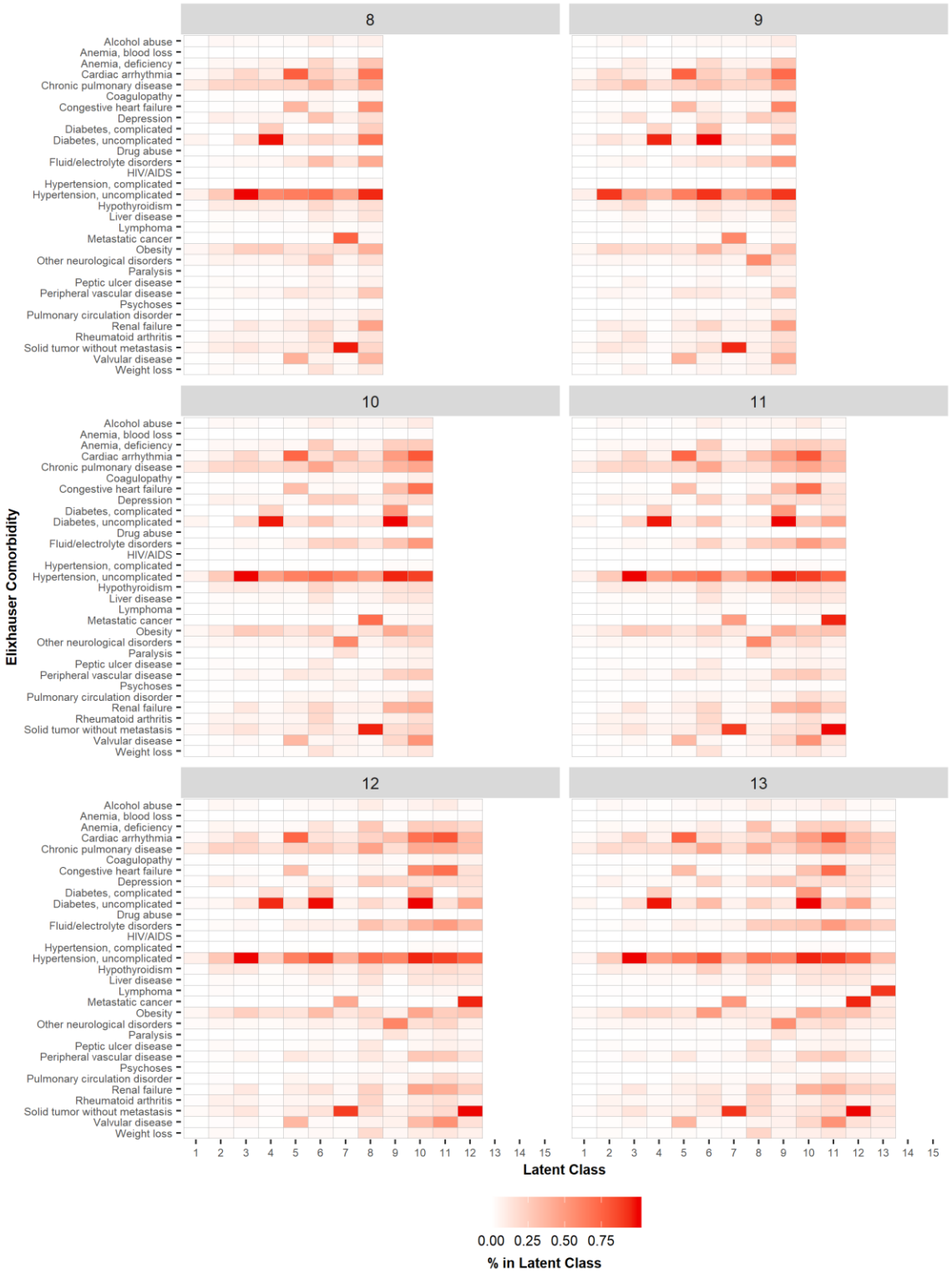
Supplementary Figure 4 (Continued): Class-Specific Item Response Probabilities for 2-15 Class Solutions, '*Participants 65 Years or Younger*' sample (n=63,032)



Supplementary Figure 5: Latent Class Analysis Model Results for 2-15 Class Solutions, '*Participants Over 65 Years*' sample (n=107,702)

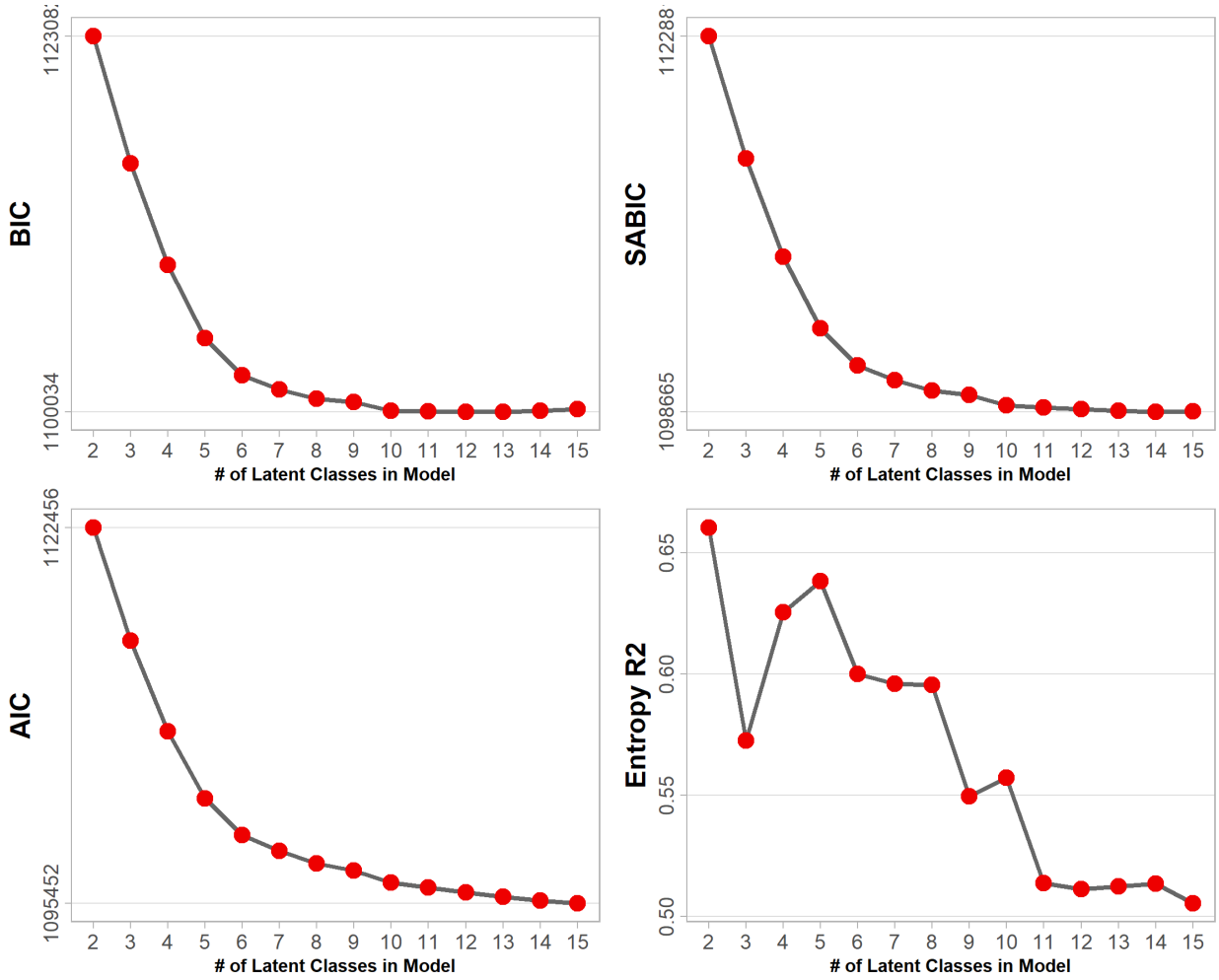


Supplementary Figure 6: Class-Specific Item Response Probabilities for 2-15 Class Solutions, '*Participants Over 65 Years*' sample (n=107,702)

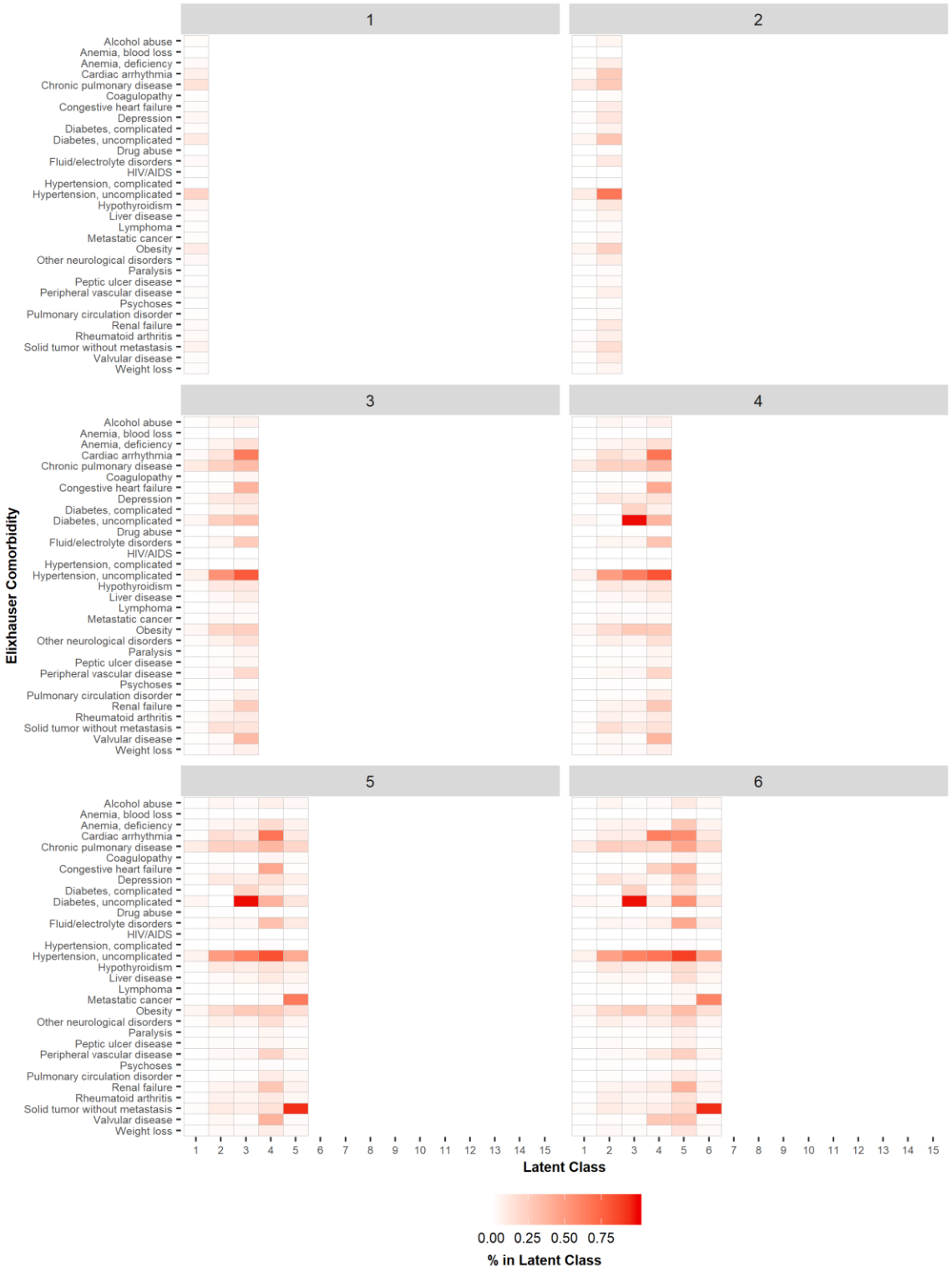


Supplementary Figure 6 (Continued): Class-Specific Item Response Probabilities for 2-15 Class Solutions, '*Participants Over 65 Years*' sample (n=107,702)

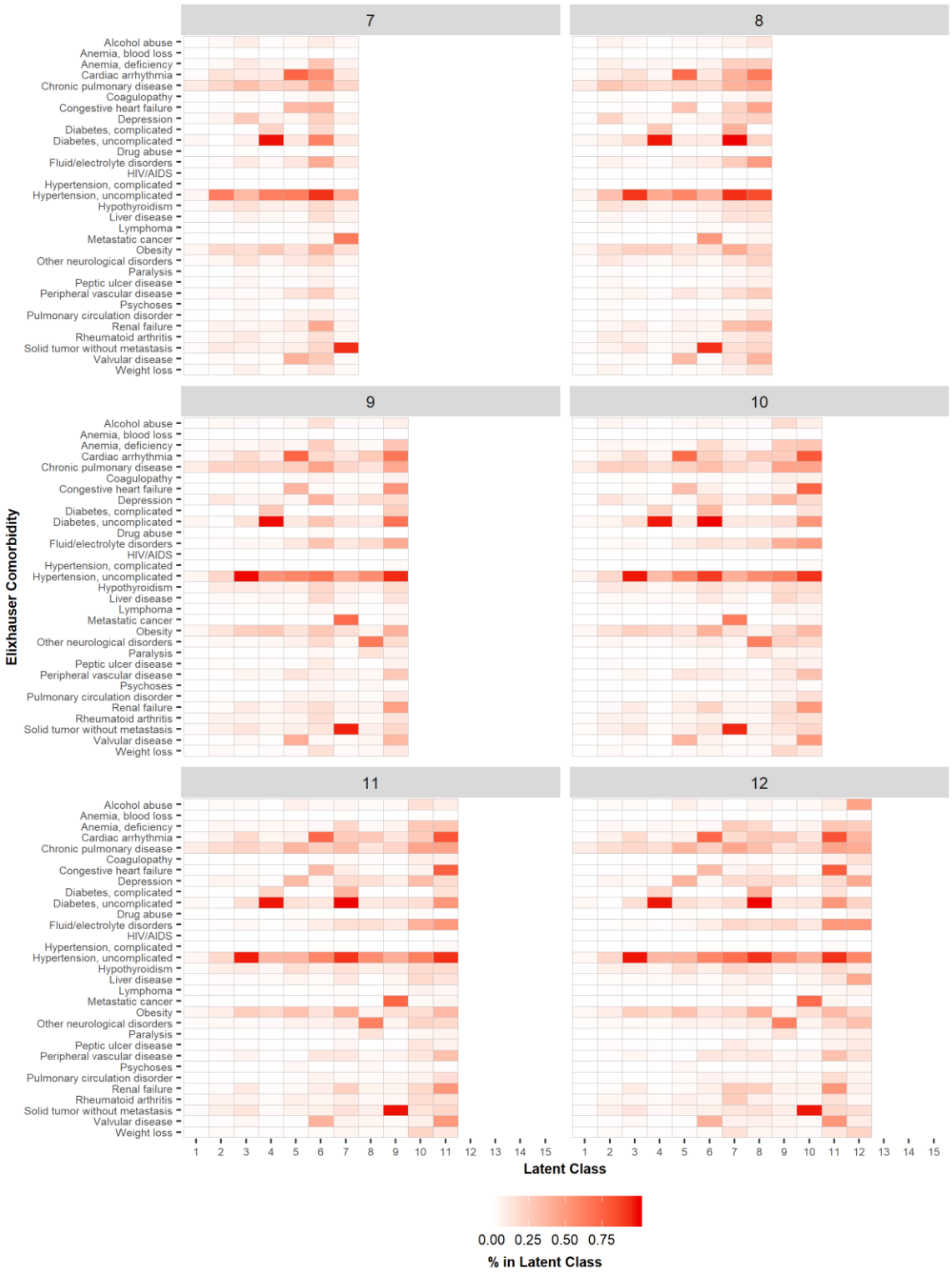
Supplementary Figure 6 (Continued): Class-Specific Item Response Probabilities for 2-15 Class Solutions, '*Participants Over 65 Years*' sample (n=107,702)



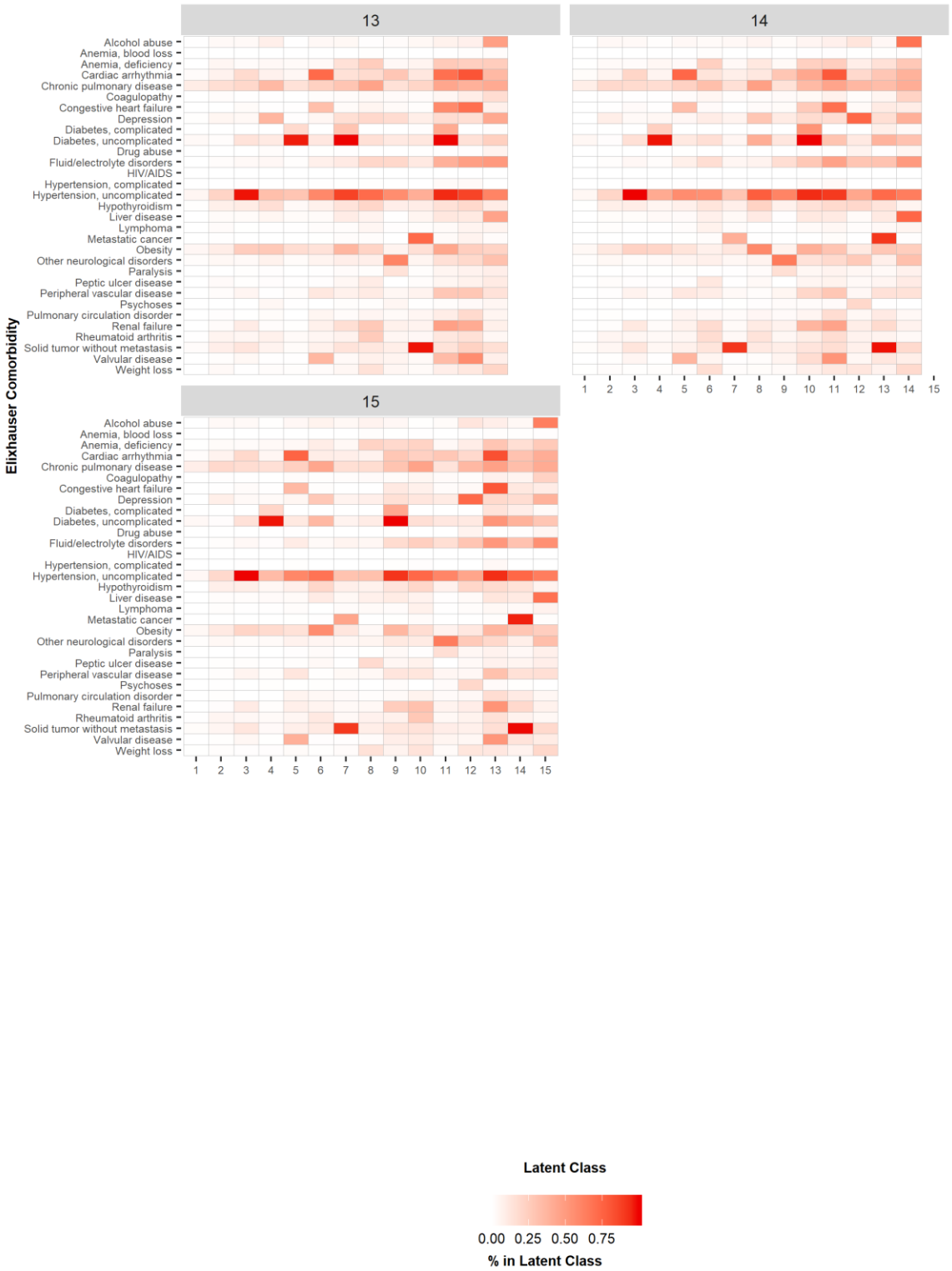
Supplementary Figure 7: Latent Class Analysis Model Results for 2-15 Class Solutions, 'Unrelated' sample (n=151,623)



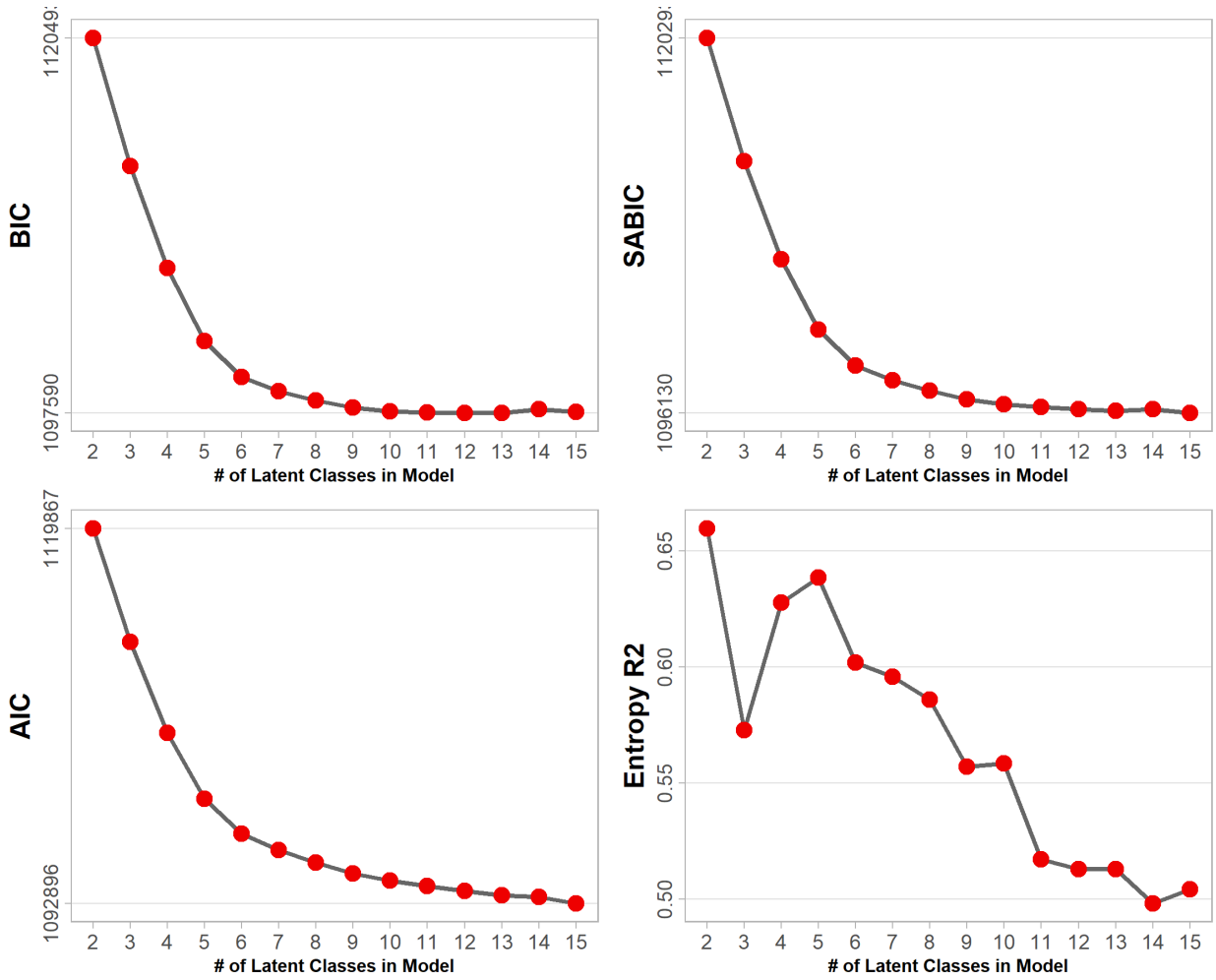
Supplementary Figure 8: Class-Specific Item Response Probabilities for 2-15 Class Solutions, 'Unrelated' sample (n=151,623)



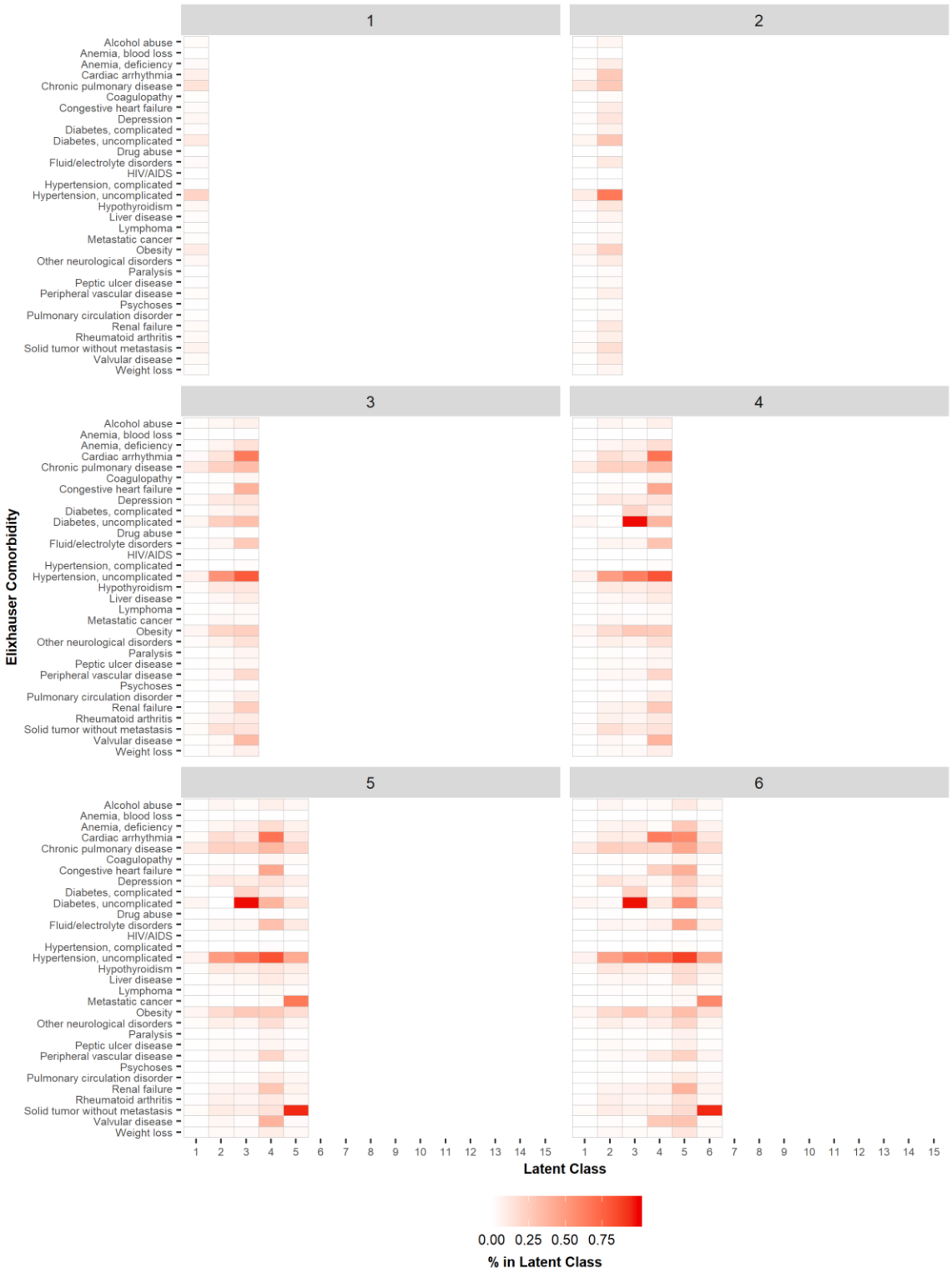
Supplementary Figure 8 (Continued): Class-Specific Item Response Probabilities for 2-15 Class Solutions, 'Unrelated' sample (n=151,623)



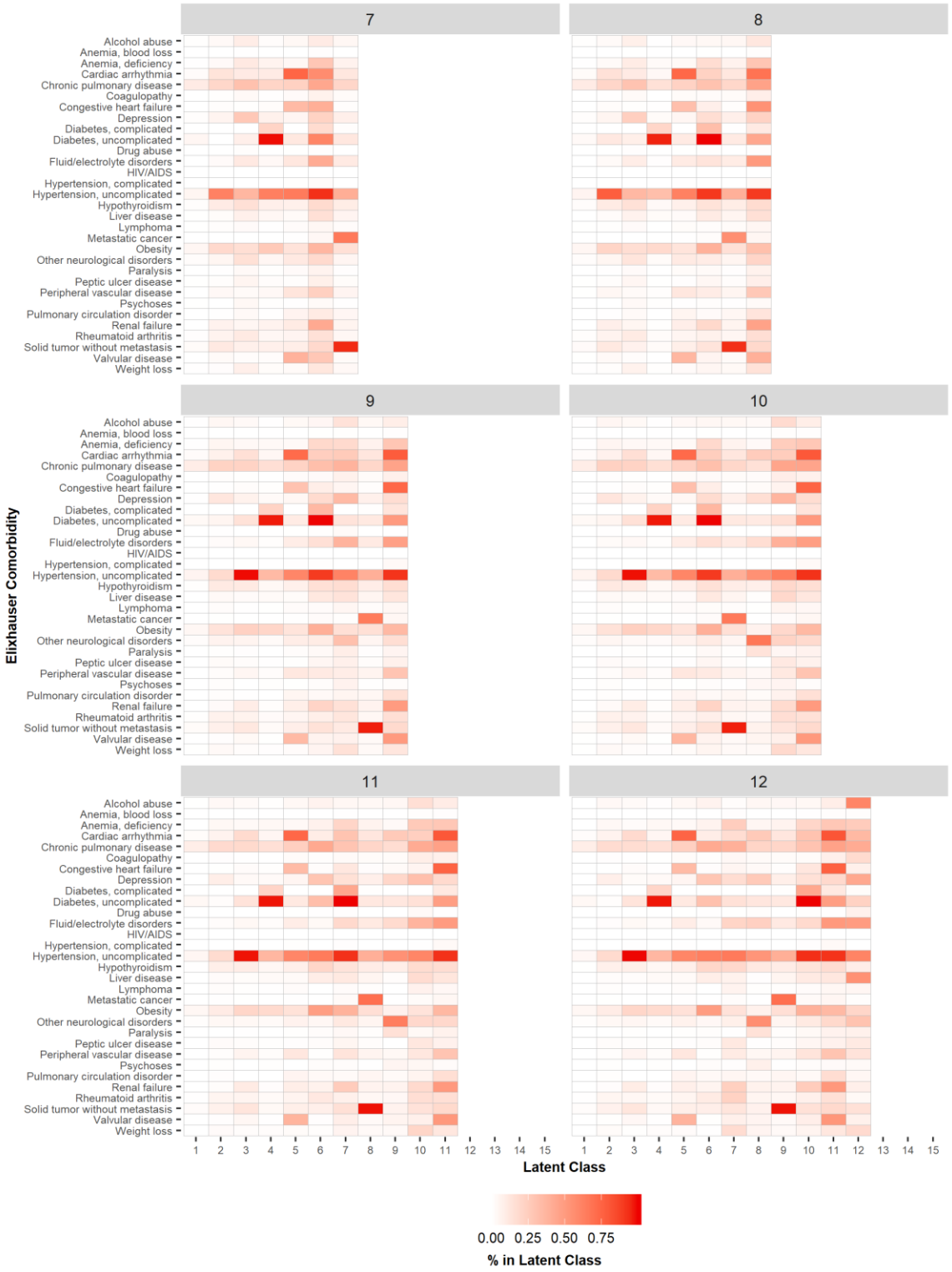
Supplementary Figure 8 (Continued): Class-Specific Item Response Probabilities for 2-15 Class Solutions, 'Unrelated' sample (n=151,623)



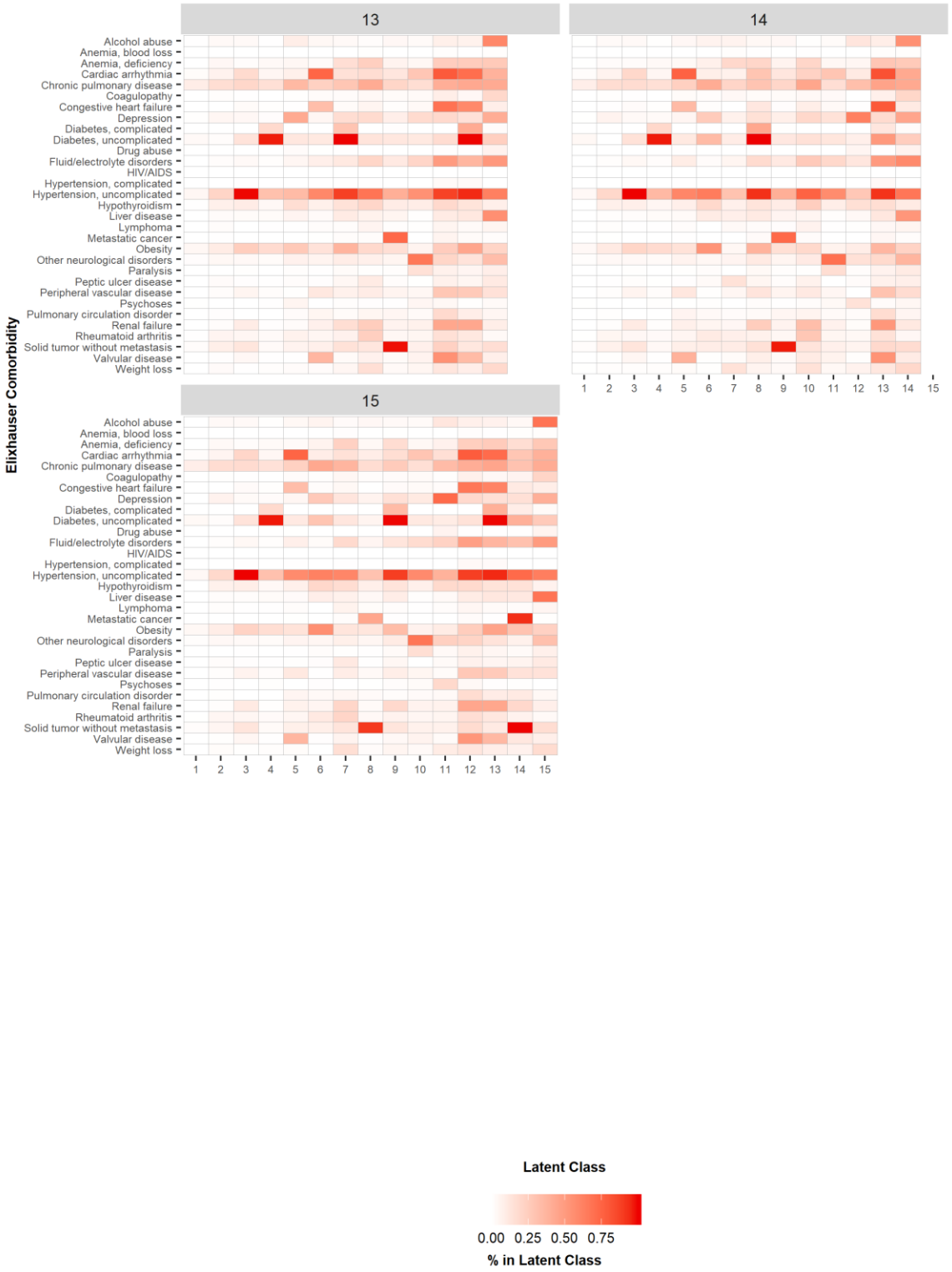
Supplementary Figure 9: Latent Class Analysis Model Results for 2-15 Class Solutions, 'Mixed kinship' sample (n=151,623)



Supplementary Figure 10: Class-Specific Item Response Probabilities for 2-15 Class Solutions, 'Mixed kinship' sample (n=151,623)



Supplementary Figure 10 (Continued): Class-Specific Item Response Probabilities for 2-15 Class Solutions, 'Mixed kinship' sample (n=151,623)



Supplementary Figure 10 (Continued): Class-Specific Item Response Probabilities for 2-15 Class Solutions, 'Mixed kinship' sample (n=151,623)