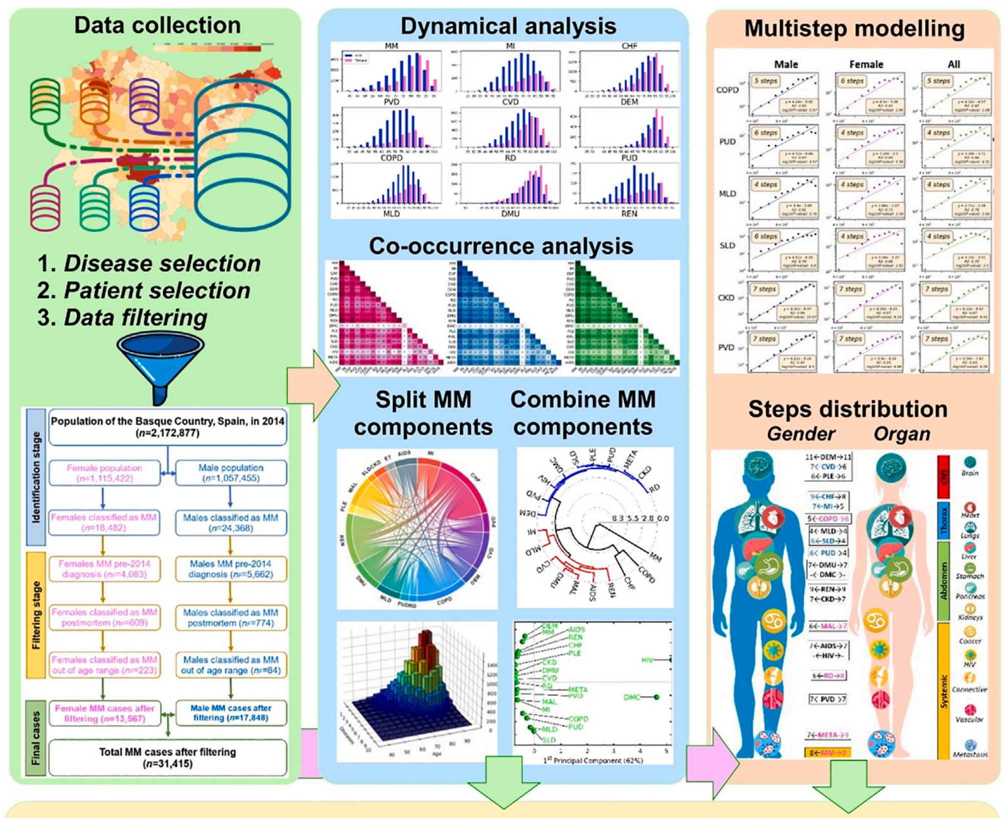
iScience

CellPress
OPEN ACCESS

Article

Chronic disease incidence explained by stepwise models and co-occurrence among them


Data collection
1. Disease selection
2. Patient selection
3. Data filtering
Identification stage
Population of the Basque Country, Spain, in 2014 (n=2,172,877)
Female population (n=1,115,422)
Male population (n=1,057,455)
Females classified as MM (n=18,482)
Males classified as MM (n=24,368)
Filtering stage
Females MM pre-2014 diagnosis (nF=4,083)
Males MM pre-2014 diagnosis (nF=5,662)
Females classified as MM postmortem (nF=609)
Males classified as MM postmortem (nF=774)
Females classified as MM out of age range (nF=223)
Males classified as MM out of age range (nF=84)
Final cases
Female MM cases after filtering (n=13,567)
Male MM cases after filtering (n=17,848)
Total MM cases after filtering (n=31,415)
Dynamical analysis
MM
MI
CHF
PVD
CVD
DEM
COPD
RD
PUD
MLD
DMU
REN
Co-occurrence analysis
Split MM components
Combine MM components
Multistep modelling
Male
Female
All
COPD
PUD
MLD
SLD
CKD
PVD
Steps distribution
Gender
Organ
Brain
Heart
Lungs
Liver
Stomach
Pancreas
Kidneys
Cancer
HIV
Connective
Vascular
Metastasis
CNS
Thorax
Abdomen
Systemic

- The Charlson Comorbidity Index (CCI) diseases delineate a complex interaction network.
- The CCI incidence-age profiles follow three dynamic categories.
- A global analysis of the CCI incidence-age profiles discriminates between low- and high-risk of dying pathologies.
- The incidence-age profiles of multimorbidity and six CCIs follow confidently multistep models.
- Multimorbidity follows a multistep model with eight steps for males and nine for females.

Mikel Arróspide Elgarresta, Daniela Gerovska, Myrian Soto-Gordoa, María L. Jauregui García, Marisa L. Merino Hernández, Marcos J. Araúzo-Bravo

mararabra@yahoo.co.uk

Highlights

Charlson Comorbidity Index (CCI) diseases delineate a complex interaction network

CCI incidence-age profiles discriminate between low- and high-risk of dying pathologies

Incidence-age profiles of multimorbidity and 6 CCIs follow confidently multistep models

Multimorbidity follows a multistep model with 8 steps for males and 9 for females

Arróspide Elgarresta et al., iScience 27, 110816
September 20, 2024 © 2024 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.isci.2024.110816

iScience

CellPress
OPEN ACCESS

Article

Chronic disease incidence explained by stepwise models and co-occurrence among them

Mikel Arróspide Elgarresta,[1] Daniela Gerovska,[1] Myrian Soto-Gordoa,[2,3] María L. Jauregui García,[2,4] Marisa L. Merino Hernández,[2,5,6] and Marcos J. Araúzo-Bravo[1,7,8,9,10,11,*]

SUMMARY

Multimorbidity (MM) is the co-occurrence of two or more chronic diseases. We provided a dynamic approach revealing the MM complexity constructing a multistep incidence-age model for all patients with MM between 2014 and 2021 in the Basque Health System, Spain. The multistep model, with eight steps for males and nine for females, is a very well-fitting representation of MM. To gain insight into the MM components, we modeled the 19 diseases used to calculate the Charlson Comorbidity Index (CCI). We observed that the CCI diseases formed a complex interaction network. Hierarchical clustering of the incidence-age profiles clustered the CCI diseases into low- and high-risk of dying pathologies. Diseases with a higher number of steps are better represented by a multistep model. Anatomically, diseases associated with the central nervous system have the highest number of steps, followed by those associated with the kidney, heart, peripheral vasulature, pancreas, joints, cerebral vasculature, lung, stomach, and liver.

INTRODUCTION

Multimorbidity (MM) is defined as the co-occurrence of two or more chronic conditions in a single patient,[1] while comorbidity refers to the presence of additional conditions, comorbidities, experienced in a patient with a specific condition of interest. Multimorbidity is a common occurrence in elderly populations who require complex care and treatment.[2] The term multimorbidity is not well established in the literature. In this study, we understand multimorbidity as the presence of at least two diseases identified by the risk stratification of the Basque Health System (BHS), Spain. The original aim of such risk stratification is to improve patients' prognosis, and therefore to identify those who were able to benefit most from integrated care optimizing resources. This process of patient identification is objective and can be transferred to other settings with universalized national systems.

A recent comprehensive meta-analysis found that the pooled prevalence of multimorbidity was 42.4% (95% CI 38.9%–46.0%) with high heterogeneity (I^2 > 99%) in a global study covering all the continents.[3] However, the stratification method used by the BHS, reported a much lower prevalence of almost 2%. Although this represents a relatively small proportion of the population, it still results in a significant consumption of resources and poses a challenge to the sustainability of the health system.[4] We use the term MM to describe patients with multiple symptomatic chronic conditions that are decompensating in nature and difficult to manage.

The impact of the condition of patients with MM is becoming increasingly important in our society due to the increase in life expectancy and the aging of the population. Therefore, it is necessary to develop novel models of personalized, predictive, preventive, participatory, and population-based care. Population risk stratification tools provide some insights for the development of new models. However, patients with MM present a wide variety of complications, due to the different pathologies they exhibit, as well as often associated issues such as aging, dependency, and social implications.

According to Skou et al.,[5] traditional statistical methods may be inadequate for stratifying patients with complex needs, particularly those with multiple chronic conditions, due to the complexity of the MM. As healthcare systems continue to experience an increase in the number of these patients, meeting their needs will require a significant amount of resources, which may threaten sustainability. To better understand the

[1]Computational Biology and Systems Biomedicine, Biogipuzkoa Health Research Institute, Calle Doctor Begiristain s/n, 20014 San Sebastian, Spain
[2]Biogipuzkoa Health Research Institute, San Sebastian-Donostia, Spain
[3]Mondragon University, Faculty of Engineering, Mondragon, Spain
[4]Tolosaldea Integrated Health Care Organization, Tolosa, Spain
[5]Bidasoa Integrated Health Care Organization, Hondarribia, Spain
[6]Research Network on Chronicity, Primary Care and Prevention and Health Promotion (RICAAPS), Kronikgune Group, Barakaldo, Spain
[7]Basque Foundation for Science, IKERBASQUE, Calle María Díaz Harokoa 3, 48013 Bilbao, Spain
[8]CIBER of Frailty and Healthy Aging (CIBERfes), 28029 Madrid, Spain
[9]Max Planck Institute for Molecular Biomedicine, Computational Biology and Bioinformatics, Röntgenstr. 20, 48149 Münster, Germany
[10]Department of Cell Biology and Histology, Faculty of Medicine and Nursing, University of Basque Country (UPV/EHU), 48940 Leioa, Spain
[11]Lead contact
*Correspondence: mararabra@yahoo.co.uk

https://doi.org/10.1016/j.isci.2024.110816

evolution of multiple chronic conditions, it may be beneficial to create mathematical models, as the optimal course of action in this complex landscape is currently unknown. These models can facilitate more informed decision-making within the healthcare system. A variety of methods have been developed to measure multimorbidity,[6] of which the Charlson Comorbidity Index (CCI) is the most extensively studied comorbidity index for predicting mortality. The Cumulative Illness Rating Scale (CIRS) assesses all relevant body systems without the use of specific diagnoses. The Index of Coexisting Disease (ICED) is a measure of disease severity and disability, while the Elixhauser Comorbidity Index (ECI) is a measure of the overall severity of comorbidities, both of which predict hospital length of stay, hospital charges, and in-hospital mortality.[7] In addition to the construction of metrics, other systems biology-oriented methods, such as the mixed graphical models (MGMs) and their integration with social network analysis techniques, have been used to study the complexity of multimorbidity.[8] Here, we propose a novel modeling approach inspired by systems biology.

A number of diseases, including cancer, are believed to follow a multistep model of pathogenesis. The multistep model of cancer pathogenesis was initially proposed by Fisher and Hollomon[9] and Armitage and Doll.[10] This model provides an explanation for the observation that cancer manifests primarily in adulthood and that the incidence of the disease increases with age. The model proposes that multiple subsequent mutations or pathogenic events are required to trigger the disease. To estimate the average number of pathogenic steps required, the authors modeled the patterns of cancer incidence with age. The number of steps (n) was estimated by calculating the slope (m) of the regression between $\log_{10}(incidence)$ and $\log_{10}(age)$ and adding 1, resulting in $n = m + 1$.

A number of researchers[11–17] have employed the original multistep model[10] to investigate the hypothesis that certain neurodegenerative diseases (NDs) may also be multistep processes. In particular, Al-Chalabi et al.[11] employed a multistep model for amyotrophic lateral sclerosis (ALS). The study revealed a linear relationship between the logarithm of incidence and the logarithm of age, indicating a six-step process. In a comprehensive multistep analysis of various NDs, Gerovska et al.[14] constructed a genealogical tree of the NDs and determined the number of steps required to initiate each ND disease. Gerovska and Araúzo-Bravo[18] showed that a wider age range of onset is associated with a reduction in the number of ND steps.

Multistep modeling may also prove useful for multimorbidity, even if the condition is not a traditional index disease in itself. Rather, multimorbidity represents a heterogeneous group of patients who may be very different from each other. The way we approach such a conundrum is to deconvolute the problem by splitting it down into more elementary components. There are multiple potential ways for identifying these elementary components — i.e., unsupervised clustering of patents or supervised clustering of patients with different types of supervision signals. Since our working definition of multimorbidity is the co-occurrence of at least two diseases, we use as elementary elements the multiple diseases associated with multimorbidity. From the point of view of data collection this is a straightforward decomposition, however, such decomposition does not produce a disjoint separation of the patients since by definition each patient suffers at least from one disease. Anyway, it offers a simple way to undertake the problem. Once the multistep models for each disease have been built, we compare such elementary models with our objective general multistage model covering all the multimorbidity patients.

Therefore, for our available data, our study explores the incidence-age multistep model as a possible explanation for all the patients with MM in the Basque Country of Spain. The study analyzes the incidence age of the 19 diseases used to construct the CCI comorbidities, with the aim of deconstructing the main components of the multimorbidity ecosystem. The data for the MM and each CCI disease are evaluated for the fit of multistep models, with the aim of gaining a better understanding of the underlying factors that contribute to the development of these conditions.

The approach used in the study of multistep modeling is advantageous for elucidating the characteristics of multimorbid patients, who are a heterogeneous population in terms of clinical entities and health outcomes. Previously, these patients have been classified and characterized according to different multimorbidity patterns. Here, the primary components of multimorbidity are divided and modeled in a parametrized approach, which is particularly suitable for characterizing the varying incidence rates of each disease. Such modeling allows for a comparison of the number of steps required to trigger each disease component of the multimorbidity and with the multimorbidity multistep model itself. This approach provides a framework for disentangling the different components of its complexity and elucidating the manner in which the distinct diseases contribute to the triggering of multimorbidity at varying rates. This perspective is complementary to other modeling techniques, such as proportional hazard models, which are more appropriate for studying associations with risk factors[19] of Cox models.[20] Moreover, the alignment of the incidence age for all diseases permits the visualization of the similarities and the clustering of different disease profiles.

To conduct our study, we obtained a comprehensive dataset comprising all patients registered with the Basque Health System (BHS) in Spain between 2014 and 2021, who had been labeled by the system as MM (Figure 1). Subsequently, we extracted a list of the 19 diseases that were used in the calculation of the scoring scheme CDMF (Claims-based, Disease-specific refinements, Matching translation to ICD10, Flexibility) for the revised Charlson Comorbidity Index (CCI). From the resulting list of 20 diseases, 19 plus MM, we performed a comprehensive analysis of the incidence rates of individual chronic diseases by age. We then studied the co-occurrence of all pairs of chronic diseases and aligned the dynamic data for all diseases to construct a matrix that allowed us to perform a global comparative analysis of all diseases. Finally, we constructed a stepwise model to explain the development of individual chronic diseases and MM and studied the diseases that fit such a model.

RESULTS

The distribution of the number of diagnoses for multimorbidity and each of the Charlson comorbidity index pathologies indicates a complex interaction network

To provide a comprehensive analysis of the incidence rates of individual chronic diseases by age we study the distribution of the number and percentages of patients for each of the CCI diseases and genders. The analysis shows that the most prevalent diseases among both males and

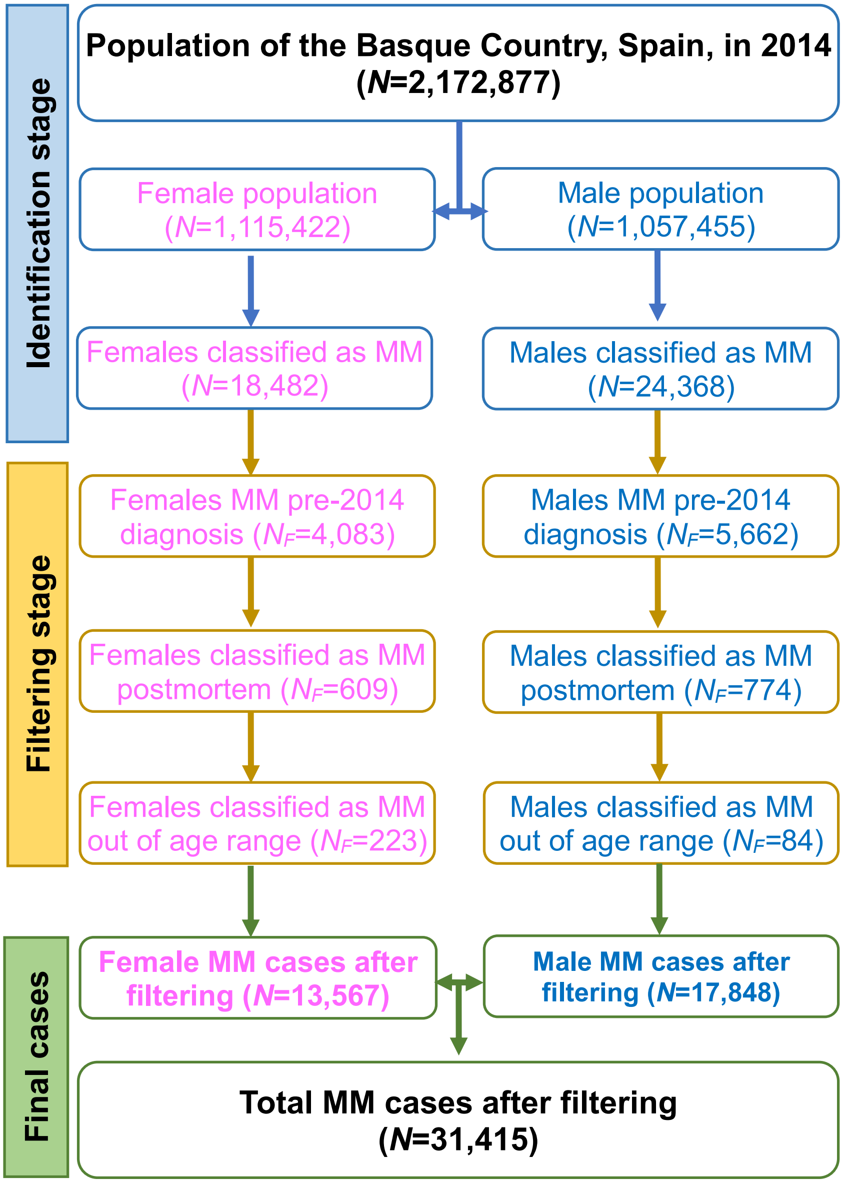
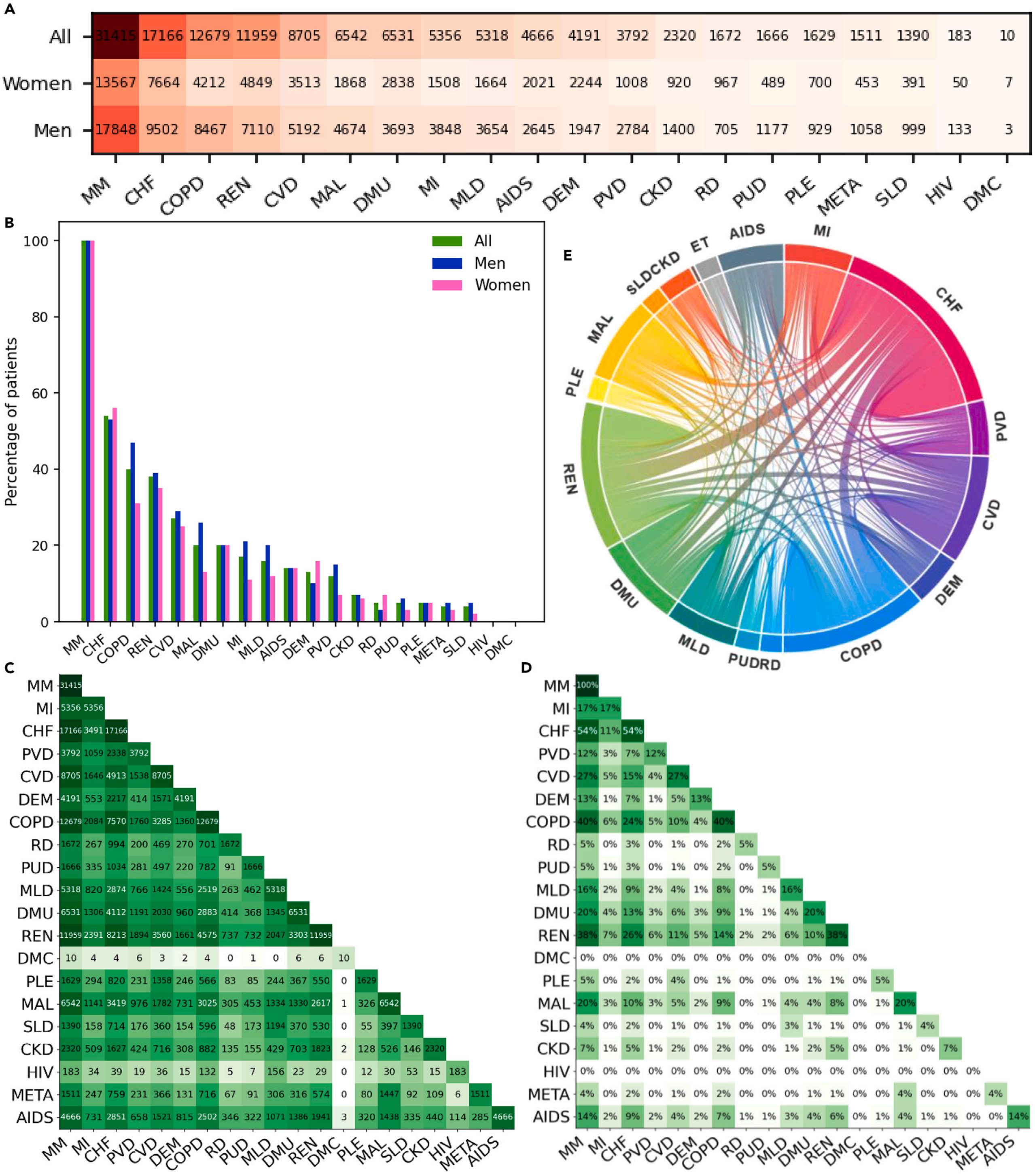

Identification stage
Filtering stage
Final cases
Population of the Basque Country, Spain, in 2014 (N=2,172,877)
Female population (N=1,115,422)
Male population (N=1,057,455)
Females classified as MM (N=18,482)
Males classified as MM (N=24,368)
Females MM pre-2014 diagnosis (NF=4,083)
Males MM pre-2014 diagnosis (NF=5,662)
Females classified as MM postmortem (NF=609)
Males classified as MM postmortem (NF=774)
Females classified as MM out of age range (NF=223)
Males classified as MM out of age range (NF=84)
Female MM cases after filtering (N=13,567)
Male MM cases after filtering (N=17,848)
Total MM cases after filtering (N=31,415)

Figure 1. Flowchart of the collection, selection, and extraction of data to calculate the multistep models of the MM and the CCI pathologies

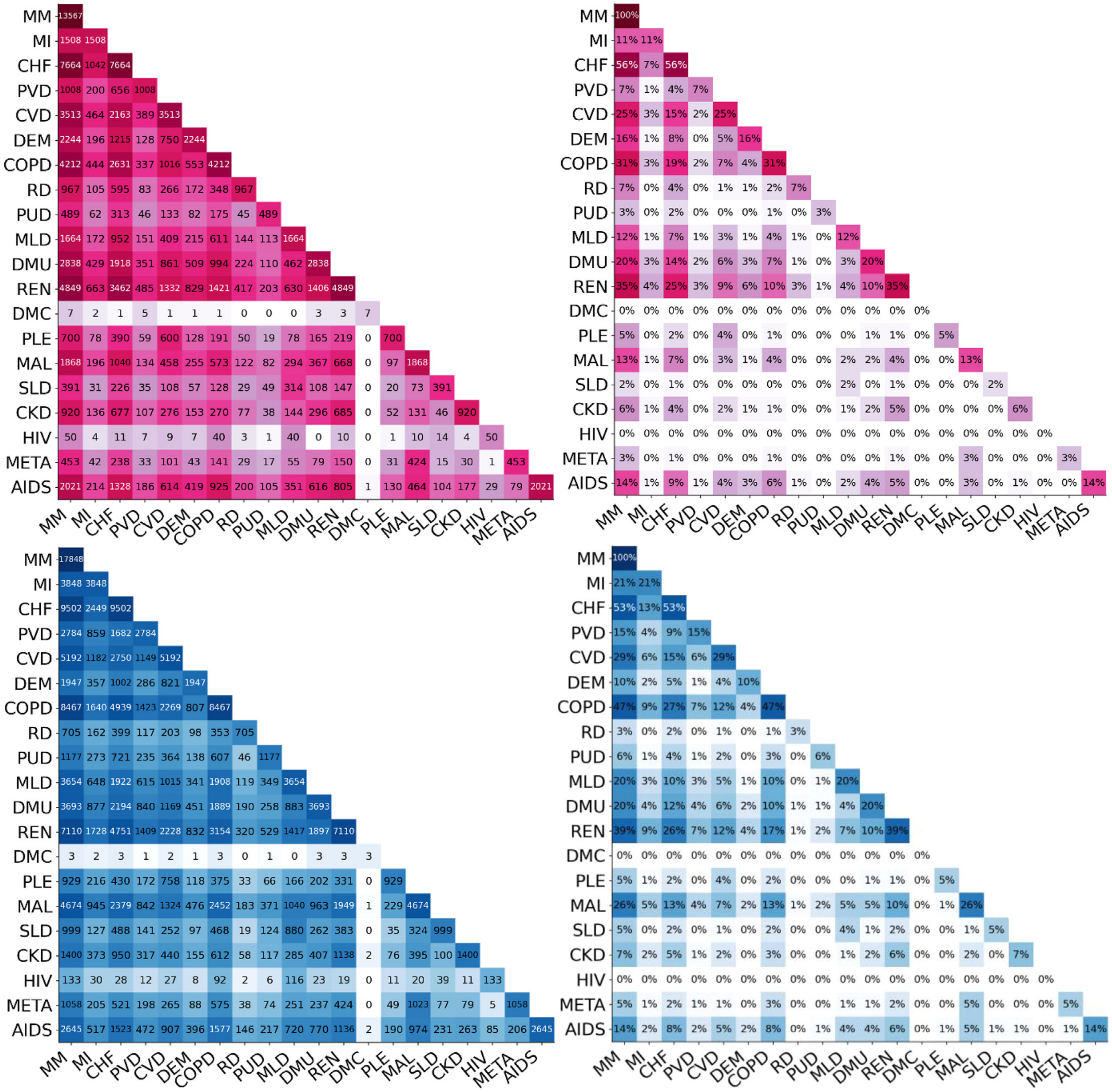
Male and female cases are shown in blue and pink, respectively; *N* and N_F denote the numbers of remaining and filtered cases in each stage respectively.

females are CHF[1] (56% of females, 53% of males), COPD[1] (31% of females, 47% of males), and REN[1] (35% of females, 39% of males), all with a CCI score of 1—the numbers in brackets represent the CCI of each disease. On the other hand, DMC[2], HIV[3], and SLD[3] are the least prevalent diseases, all with a CCI score higher than 1 (Figures 2A and 2B). It is noteworthy that males have a higher incidence of almost all CCI diseases, except for CHF[1], DEM[1], and RD[1], where females have a higher incidence. However, both genders have the same incidence for DMU[1], AIDS[6], and PLE[2].

To gain insight into the co-occurrence of diseases associated with CCI, an analysis of the frequency pairs of all possible comorbidities was conducted. The analysis of diseases that occur simultaneously in patients revealed two distinct types of diseases, as shown in Figures 2C and 2D. The first category includes diseases that are frequently concomitant with other diseases with a co-occurrence rate of at least 10%. These are primarily observed in the upper triangular sub-array of the heatmaps mainly constrained by DMC. This group includes CHF[1], CVD[1], COPD[1], DMU[1], REN[1], and MAL[2]. Except for MAL, all these diseases have CCI score of 1. The second category comprises diseases that are less frequently associated with other diseases, with a co-occurrence rate of less than 10%. This category includes MI[1], PVD [1], DEM[1], RD[1], PUD[1], MLD[1], DMC[2], PLE[2], SLD[3], CKD[3], HIV[3], META [6], and AIDS[6], with more variable CCI scores. As illustrated in Figures 2C and 2D, the data demonstrate that the most prevalent



(figure continued on next page)



	MM	MI	CHF	PVD	CVD	DEM	COPD	RD	PUD	MLD	DMU	REN	DMC	PLE	MAL	SLD	CKD	HIV	META	AIDS
MM	13567																			
MI	1508	1508																		
CHF	7664	1042	7664																	
PVD	1008	200	656	1008																
CVD	3513	464	2163	389	3513															
DEM	2244	196	1215	128	750	2244														
COPD	4212	444	2631	337	1016	553	4212													
RD	967	105	595	83	266	172	348	967												
PUD	489	62	313	46	133	82	175	45	489											
MLD	1664	172	952	151	409	215	611	144	113	1664										
DMU	2838	429	1918	351	861	509	994	224	110	462	2838									
REN	4849	663	3462	485	1332	829	1421	417	203	630	1406	4849								
DMC	7	2	1	5	1	1	1	0	0	0	3	3	7							
PLE	700	78	390	59	600	128	191	50	19	78	165	219	0	700						
MAL	1868	196	1040	134	458	255	573	122	82	294	367	668	0	97	1868					
SLD	391	31	226	35	108	57	128	29	49	314	108	147	0	20	73	391				
CKD	920	136	677	107	276	153	270	77	38	144	296	685	0	52	131	46	920			
HIV	50	4	11	7	9	7	40	3	1	40	0	10	0	1	10	14	4	50		
META	453	42	238	33	101	43	141	29	17	55	79	150	0	31	424	15	30	1	453	
AIDS	2021	214	1328	186	614	419	925	200	105	351	616	805	1	130	464	104	177	29	79	2021

	MM	MI	CHF	PVD	CVD	DEM	COPD	RD	PUD	MLD	DMU	REN	DMC	PLE	MAL	SLD	CKD	HIV	META	AIDS
MM	100%																			
MI	11%	11%																		
CHF	56%	7%	56%																	
PVD	7%	1%	4%	7%																
CVD	25%	3%	15%	2%	25%															
DEM	16%	1%	8%	0%	5%	16%														
COPD	31%	3%	19%	2%	7%	4%	31%													
RD	7%	0%	4%	0%	1%	1%	2%	7%												
PUD	3%	0%	2%	0%	0%	0%	1%	0%	3%											
MLD	12%	1%	7%	1%	3%	1%	4%	1%	0%	12%										
DMU	20%	3%	14%	2%	6%	3%	7%	1%	0%	3%	20%									
REN	35%	4%	25%	3%	9%	6%	10%	3%	1%	4%	10%	35%								
DMC	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%							
PLE	5%	0%	2%	0%	4%	0%	1%	0%	0%	0%	1%	1%	0%	5%						
MAL	13%	1%	7%	0%	3%	1%	4%	0%	0%	2%	2%	4%	0%	0%	13%					
SLD	2%	0%	1%	0%	0%	0%	0%	0%	0%	2%	0%	1%	0%	0%	0%	2%				
CKD	6%	1%	4%	0%	2%	1%	1%	0%	0%	1%	2%	5%	0%	0%	0%	0%	6%			
HIV	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%		
META	3%	0%	1%	0%	0%	0%	1%	0%	0%	0%	0%	1%	0%	0%	3%	0%	0%	0%	3%	
AIDS	14%	1%	9%	1%	4%	3%	6%	1%	0%	2%	4%	5%	0%	0%	3%	0%	1%	0%	0%	14%

	MM	MI	CHF	PVD	CVD	DEM	COPD	RD	PUD	MLD	DMU	REN	DMC	PLE	MAL	SLD	CKD	HIV	META	AIDS
MM	17848																			
MI	3848	3848																		
CHF	9502	2449	9502																	
PVD	2784	859	1682	2784																
CVD	5192	1182	2750	1149	5192															
DEM	1947	357	1002	286	821	1947														
COPD	8467	1640	4939	1423	2269	807	8467													
RD	705	162	399	117	203	98	353	705												
PUD	1177	273	721	235	364	138	607	46	1177											
MLD	3654	648	1922	615	1015	341	1908	119	349	3654										
DMU	3693	877	2194	840	1169	451	1889	190	258	883	3693									
REN	7110	1728	4751	1409	2228	832	3154	320	529	1417	1897	7110								
DMC	3	2	3	1	2	1	3	0	1	0	3	3	3							
PLE	929	216	430	172	758	118	375	33	66	166	202	331	0	929						
MAL	4674	945	2379	842	1324	476	2452	183	371	1040	963	1949	1	229	4674					
SLD	999	127	488	141	252	97	468	19	124	880	262	383	0	35	324	999				
CKD	1400	373	950	317	440	155	612	58	117	285	407	1138	2	76	395	100	1400			
HIV	133	30	28	12	27	8	92	2	6	116	23	19	0	11	20	39	11	133		
META	1058	205	521	198	265	88	575	38	74	251	237	424	0	49	1023	77	79	5	1058	
AIDS	2645	517	1523	472	907	396	1577	146	217	720	770	1136	2	190	974	231	263	85	206	2645

	MM	MI	CHF	PVD	CVD	DEM	COPD	RD	PUD	MLD	DMU	REN	DMC	PLE	MAL	SLD	CKD	HIV	META	AIDS
MM	100%																			
MI	21%	21%																		
CHF	53%	13%	53%																	
PVD	15%	4%	9%	15%																
CVD	29%	6%	15%	6%	29%															
DEM	10%	2%	5%	1%	4%	10%														
COPD	47%	9%	27%	7%	12%	4%	47%													
RD	3%	0%	2%	0%	1%	0%	1%	3%												
PUD	6%	1%	4%	1%	2%	0%	3%	0%	6%											
MLD	20%	3%	10%	3%	5%	1%	10%	0%	1%	20%										
DMU	20%	4%	12%	4%	6%	2%	10%	1%	1%	4%	20%									
REN	39%	9%	26%	7%	12%	4%	17%	1%	2%	7%	10%	39%								
DMC	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%							
PLE	5%	1%	2%	0%	4%	0%	2%	0%	0%	0%	1%	1%	0%	5%						
MAL	26%	5%	13%	4%	7%	2%	13%	1%	2%	5%	5%	10%	0%	1%	26%					
SLD	5%	0%	2%	0%	1%	0%	2%	0%	0%	4%	1%	2%	0%	0%	1%	5%				
CKD	7%	2%	5%	1%	2%	0%	3%	0%	0%	1%	2%	6%	0%	0%	2%	0%	7%			
HIV	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%		
META	5%	1%	2%	1%	1%	0%	3%	0%	0%	1%	1%	2%	0%	0%	5%	0%	0%	0%	5%	
AIDS	14%	2%	8%	2%	5%	2%	8%	0%	1%	4%	4%	6%	0%	1%	5%	1%	1%	0%	1%	14%

Figure 2. Distribution of the number and percentage of patients for each of the CCI diseases and sex

(A) Distribution of the number of patients.

(B) Distribution of the percentage of patients.

(C) Number of patients diagnosed with two CCI diseases.

(D) Percentage of patients diagnosed with two CCI diseases. For panels C and D, the higher the percentage, the darker the color of the heatmap cells. Green, pink, and blue cell colors correspond to all the sex combined data, female, and male, respectively.

(E) Chord diagram of the correlation between diseases. The length of each arc of the circle perimeter is proportional to the number of patients with the corresponding disease.

combinations of CCI pathologies are REN-CHF (26%), COPD-CHF (24%), CVD-CHF (15%), MI-CHF (11%), and MAL-CHF (10%). This suggests a robust correlation between congestive heart failure and other diseases. Furthermore, the cord diagram in Figure 2E illustrates a complex interaction network among the majority of the CCI diseases, with the exception of HIV and DMC, which have low incidence among the MMs.

The results demonstrate that, despite the definition of multimorbidity being the co-occurrence of two or more chronic diseases in an individual, multimorbidity encompasses a multitude of combinations of more than two diseases, occurring in various ways and forming a complex network of interactions.

The analysis of multimorbidity of the Charlson comorbidity index pathologies identifies three distinct dynamic types

To reveal the temporary trajectory of diseases and compare the different times of diagnosis, we studied the evolution of the diagnosis of the various CCI diseases in relation to the age of the patients of both sexes (Figure 3). We observed that for the majority of CCI diseases, cases exhibit a similar trend of slow growth until they reach a maximum between the ages of 60 and 85 years, depending on the disease, followed by a rapid decrease. However, HIV represents an exception to this trend, with diagnoses increasing rapidly to a maximum at approximately 35 years of age and then decreasing slowly. The majority of diseases are typically diagnosed in individuals between the ages of 40 and 90. However, there is an exception to this trend, namely DEM, which is usually diagnosed in individuals between the ages of 55 and 95 years. We observed that males tend to be diagnosed more frequently, particularly at younger ages. Conversely, due to their longer life expectancy, females are diagnosed with more diseases after the age of 90, with the exception of HIV. With regard to PVD, the number of cases is significantly higher in males until the age of 85, after which the number of cases in females reaches an equal number to that of males. With regard to DMC, due to the limited number of patients, it is not possible to reliably observe any trends.

In general, we can distinguish three types of dynamics. The first category encompasses old-age-related diseases, such as DEM, RD, and AIDS. These conditions typically manifest at advanced ages, with the median age of onset being 80 years. The second category comprises middle-aged diseases such as MI (in males), MLD, MAL, and SLD. These diseases exhibit a nearly Gaussian distribution of incidence, with a median incidence occurring around 65 years of age. The third category comprises young-age diseases, such as HIV. These diseases require a relatively short time to reach high incidences, with a median incidence occurring around 35 years of age.

A three-dimensional bar plot is presented to illustrate the distribution of patients based on the number of diagnosed CCI diseases and the age at diagnosis. As illustrated in Figure 4C, the prevalence of comorbidity is notably lower below the age of 50 and progressively increases with age, reaching a prevalence of five simultaneous diseases by the age of 85. The bar plot in Figure 4B indicates a slightly greater dispersion among males than females in Figure 4A. This may indicate that males experience a greater number of simultaneous comorbidities that begin at an earlier age. This study illustrated that the various CCI diseases have distinct dynamics. To quantify these dynamics, we will employ multistage modeling.

A global analysis of the incidence-age profiles of Charlson comorbidity index pathologies discriminates between low- and high-risk of dying pathologies

The preceding analysis and the multistep modeling study were conducted on each disease independently. Nevertheless, it would be advantageous to incorporate the multiple diseases that contribute to MM within a unified framework. Such an approach would facilitate an integrative global analysis based on a hierarchical structure of diseases and principal component analysis (PCA) of the vectors that describe the incidence dynamics across ages for each disease. Prior to undertaking such an analysis, it is first necessary to ensure that all the trajectories are aligned and encompass the same age ranges.

The hierarchical clustering for females (Figure 5D) reveals the existence of three primary disease groups, in addition to the MM, which encompasses all patients with multimorbidity. There is a small group that comprises the patients with COPD[1] and CHF[1], both with CCI scores of 1. Another group comprises REN[1], MLD[1], CVD[1], AIDS[6], MAL[2], DMU[1], and MI[1], also with a CCI of 1, except for AIDS and MAL. The last group includes DEM[1], HIV[3], DMC[2], RD[1], META[6], PLE[2], SLD[3], PUD[1], CKD[3], and PVD[1]. This third group has more variable and generally higher CCI scores.

The hierarchical clustering for males (Figure 5E) exhibits a similar pattern to that observed for females. In addition to MM, the analysis reveals the presence of three major disease clusters. The first group includes REN[1], CHF[1], COPD[1], MLD[1], and MI[1], where all diseases have a CCI score of 1. The second one includes MAL[2], CVD[1], AIDS[6], DMU[1], and PVD[1]. The third group includes DEM[1], SLD[3], PLE[2], PUD[1], HIV[3], DMC[2], CKD[3], META[6] and RD[1]. The latter exhibited more variable and elevated CCI scores than the other groups.

After merging the data for both sexes, the hierarchical clustering of the incidence-age profiles of the CCI pathologies (Figure 5F) reveals three primary clusters, apart from the MM. One small group is composed of COPD[1] and CHF[1], both with CCI scores of 1. Another group is composed of REN[1], AIDS[6], MAL[2], DMU[1], CVD[1], MLD[1] and MI[1], while the third one is composed of DEM[1], PVD[1], HIV[3], DMC[2], SLD[3], PLE[2], PUD[1], META[6], CKD[3], and RD[1]. Again, the latter group exhibited more variable and elevated CCI scores than the other two groups.

The PCAs for the three population groups—female (Figure 5A), male (Figure 5B), and all (Figure 5C)—show a comparable dispersion of diseases. We observed that in the three PCAs, the majority of diseases were distributed approximately as an arc with similar 1st Principal Component (PC) values, with the exceptions of HIV and DMC, which were positioned outside this arc. The following section will demonstrate that HIV and DMC are diseases that do not align with a multistep model. It is noteworthy that a comparable phenomenon was observed in our previous work on the modeling of multiple neurodegenerative diseases (NDs) with multistep models[18]. Similarly, multiple sclerosis was also observed to deviate from a multistep model and was also situated outside the arch of the NDs that follow a multistep model in the PCA.

In the PCA arch of diseases, five clusters can be observed, arranged from top to bottom in the 2nd principal component (PC2). One cluster is formed by MM and DEM[1], while another is formed by AIDS[6], REN[1], CHF[1], and PLE[2], both of which exhibit a positive PC2. A cluster is

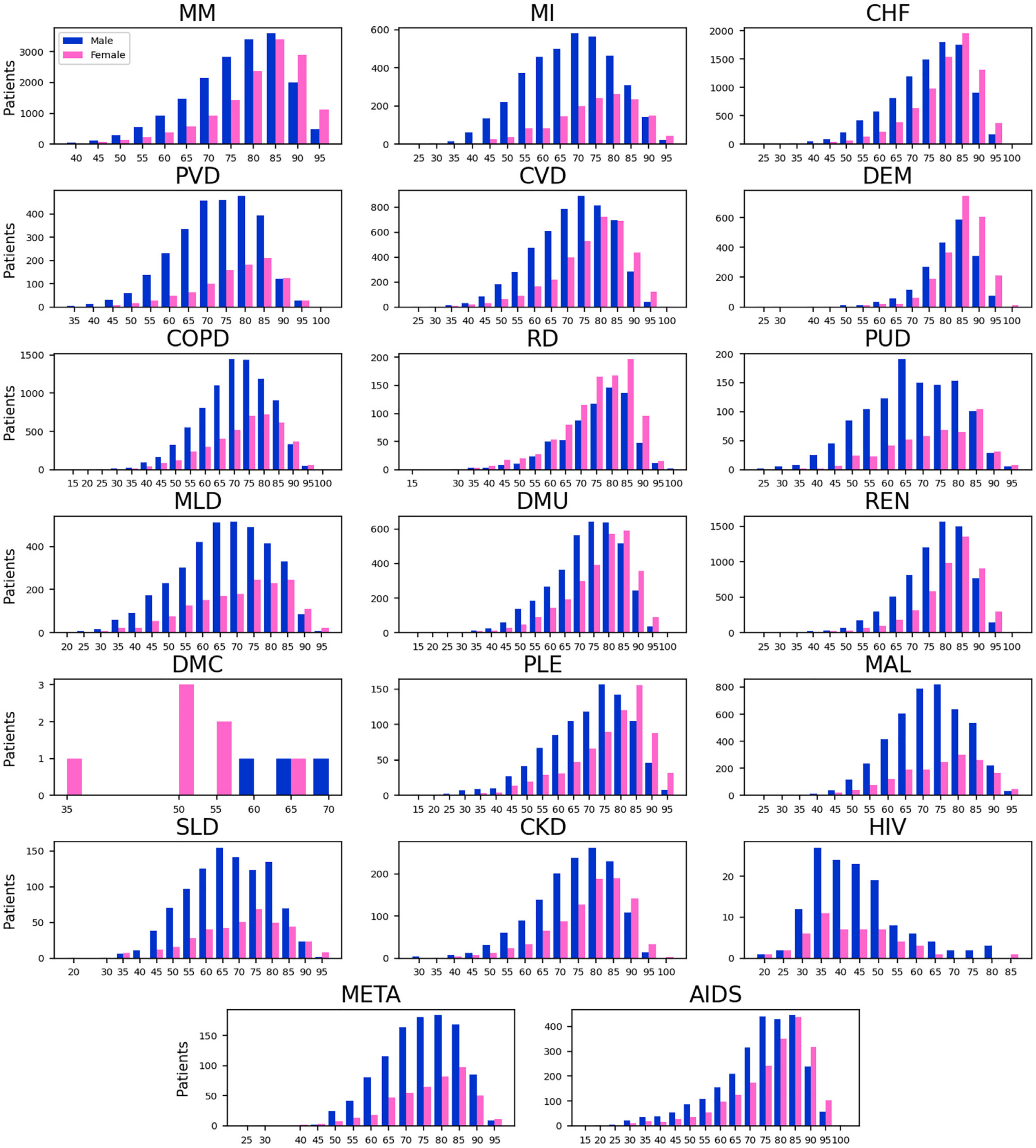


Figure 3. Bar plots of the distribution of the age at diagnosis of the different CCI diseases stratified by sex

Male and female cases are shown in blue and pink, respectively.

situated in the center, comprising CKD[3], DMU[1], PLE[2], PVD[1], MAL[2], MI[1], and META[6]. The remaining two clusters are characterized by a negative PC2, COPD[1] and PUD[1], and MLD[1] and SLD[3] (liver diseases).

With regard to the percentage of variance explained by the PCs, the sum of the first and second PCs explains 87% of the variance for the female case, and 86% for the male and combined cases. These values exceed 85%, thereby indicating that the data variation is adequately explained by these two components.

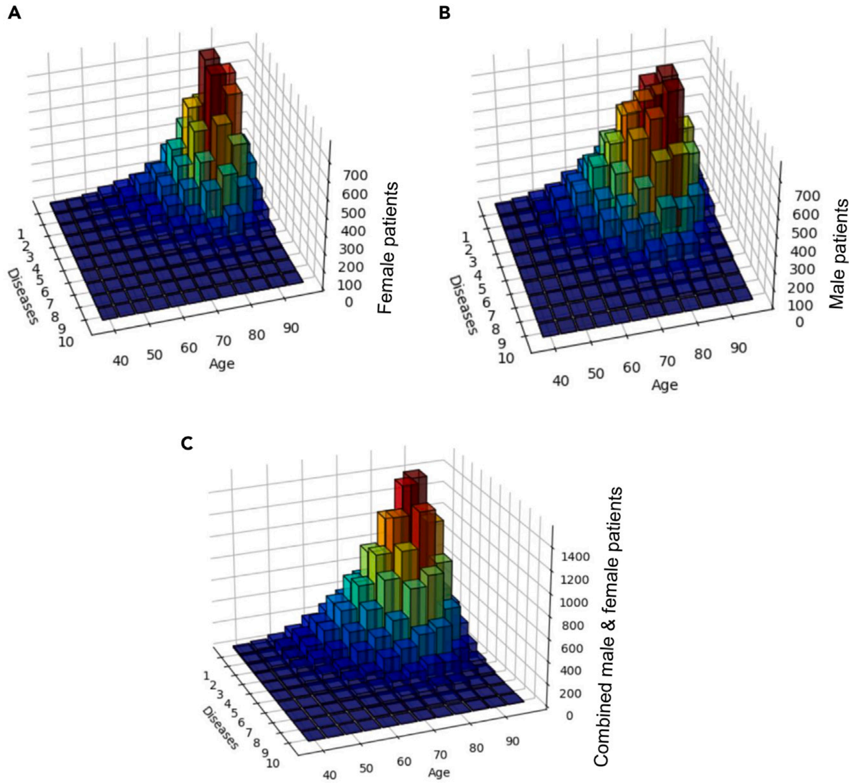


Figure 4. Tridimensional bar plots of the patients distributed by the number of CCI diseases diagnosed and the age at diagnosis

(A) Female.

(B) Male.

(C) Combined male and female data.

The age-stratified incidence profiles of multimorbidity and six Charlson comorbidity indexes follow confidently multistep models

We used a multistep modeling approach for the analysis of the age-stratified incidence profiles of all the CCI diseases, stratified by sex (Figure 6). We assessed the fit of the multistep model by measuring the quality of the regression model using R^2 and p-value metrics. The principal characteristics of the multistep models of incidence age are presented in Table 1.

We used the quality of the fit to categorize three distinct disease types: i) Confident multistep diseases, defined as those with a -log$_{10}$(p-value$_{\text{All}}$) $\geq$ 8. Under this condition the best fit diseases are CHF (-log$_{10}$(p-value$_{\text{All}}$) = 10.85), AIDS (-log$_{10}$(p-value$_{\text{All}}$) = 10.57), REN (-log$_{10}$(p-value$_{\text{All}}$) = 10.28), PLE (-log$_{10}$(p-value$_{\text{All}}$) = 9.28), DEM (-log$_{10}$(p-value$_{\text{All}}$) = 8.71), CKD (-log$_{10}$(p-value$_{\text{All}}$) = 8.52), and DMU (-log$_{10}$(p-value$_{\text{All}}$) = 8.15). These diseases are typically associated with R^2 > 0.98. ii) Possible multistep diseases, i.e., those with 8 < -log$_{10}$(p-value$_{\text{All}}$) $\geq$ 5, such as CVD (-log$_{10}$(p-value$_{\text{All}}$) = 7.75), RD (-log$_{10}$(p-value$_{\text{All}}$) = 7.53), PVD (-log$_{10}$(p-value$_{\text{All}}$) = 6.4, META (-log$_{10}$(p-value$_{\text{All}}$) = 5.73), MI (-log$_{10}$(p-value$_{\text{All}}$) = 7.66), MAL (-log$_{10}$(p-value$_{\text{All}}$) = 5.6), and COPD (-log$_{10}$(p-value$_{\text{All}}$) = 5.08). iii) Improbable multistep diseases, i.e., those with 5 < -log$_{10}$(p-value$_{\text{All}}$) $\geq$ 3, such as PUD (-log$_{10}$(p-value$_{\text{All}}$ = 4.51), MLD (-log$_{10}$(p-value$_{\text{All}}$ = 3.89), and SLD (-log$_{10}$(p-value$_{\text{All}}$) = 3.4). The last two ones cover the full range of liver disease severity, from mild to severe passing from moderate. iv) Non-multistep diseases, which are those with -log$_{10}$(p-value$_{\text{All}}$) < 3, such as HIV (-log$_{10}$(p-value$_{\text{All}}$) = 2.57), and DMC (-log$_{10}$(p-value$_{\text{All}}$) = 0.82), in which cases the model fails to an extent, resulting in negative number of steps. It is noteworthy that the MM exhibits a very good fit (-log$_{10}$(p-value$_{\text{All}}$) = 15.13).

The MM has eight steps for males and nine for females (see Figure 6, 7th row of right panel, and 3rd row of Table 1). The disease with the highest number of steps is DEM (11 steps for both males and females), followed by REN (nine steps for both males and females). The diseases with the fewest steps are MLD (4 steps for both males and females), and SLD and PLE (4 steps for females and 6 steps for males).

To contextualize these numbers of steps, it is essential to recognize that they represent the number of events that must occur in order to trigger a disease that follows a multistep model. This concept was previously proposed in the field of oncology by Fisher and Hollomon[9] and Armitage and Doll,[10] who identified that cancers require approximately five steps. In our modeling of neurodegenerative diseases (NDs),[14,18] we found that multistep models for NDs range from 2 to 12 steps. Furthermore, we identified three categories of NDs based on their step count: those with a low number of steps ($n \leq 3$), those with an intermediate number of steps ($3 < n \leq 7$), and those with a high number of steps ($7 < n$). In the present study, we found that the CCI diseases, which follow a multistep model, have a step count ranging from 4 to 11, indicating a number of steps that fall in the intermediate to high range.

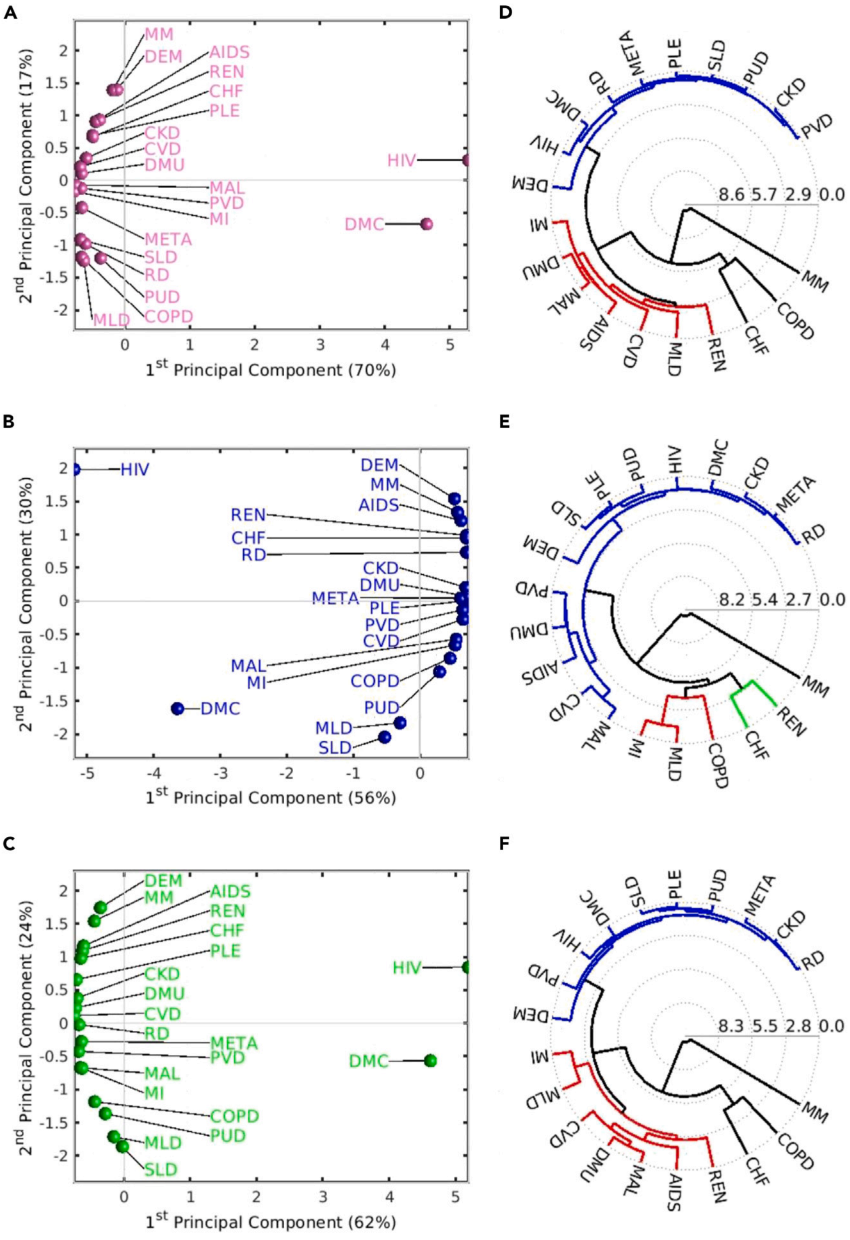


Figure 5. Global analysis of the incidence-age profiles of the CCI pathologies

Bidimensional principal component analysis (PCA) for female (A), male (B), and all sex combined data (C). Circular hierarchical clustering for female (D), male (E), and all, sex combined data (F). The metric of the clustering is the standardized Euclidean metric.

Correlations of fitness values and regression parameters

In order to gain insights into the results of the multistep regression model, we studied the possible correlation between the fitness values of the regression model and the estimated regression parameters. We observed that, in general, an increased number of steps corresponded to a superior fit of the age-incidence data to a multistep model (Figure 7A). The number of steps is proportional to the absolute value of the intercept $c = \log_{10}(u_1 \cdot u_2 \cdot u_3 \cdot \ \ldots \ u_{n-1} \cdot u_n) = \log_{10}(u)$, where $u = \exp(c)$ represents the background risk u of all steps (Figure 7B). Thus, as the intercept c is negative in all the cases, a higher number of steps is associated with a lower background risk u. As anticipated, there is a positive correlation between the two metrics of the fitness of the multistep regression model, namely the R^2 and the $\log_{10}$ of the p-value. This is due to the fact that both metrics estimate the fit of the multistep regression model to the data in different ways (Figure 7C). Moreover, a negative correlation was observed between the two estimated parameters of the regression model, the slope m and the intercept (Figure 7D). No correlation was observed between the number of steps and the CCI, indicating that the number of steps may not be associated with the disease severity.

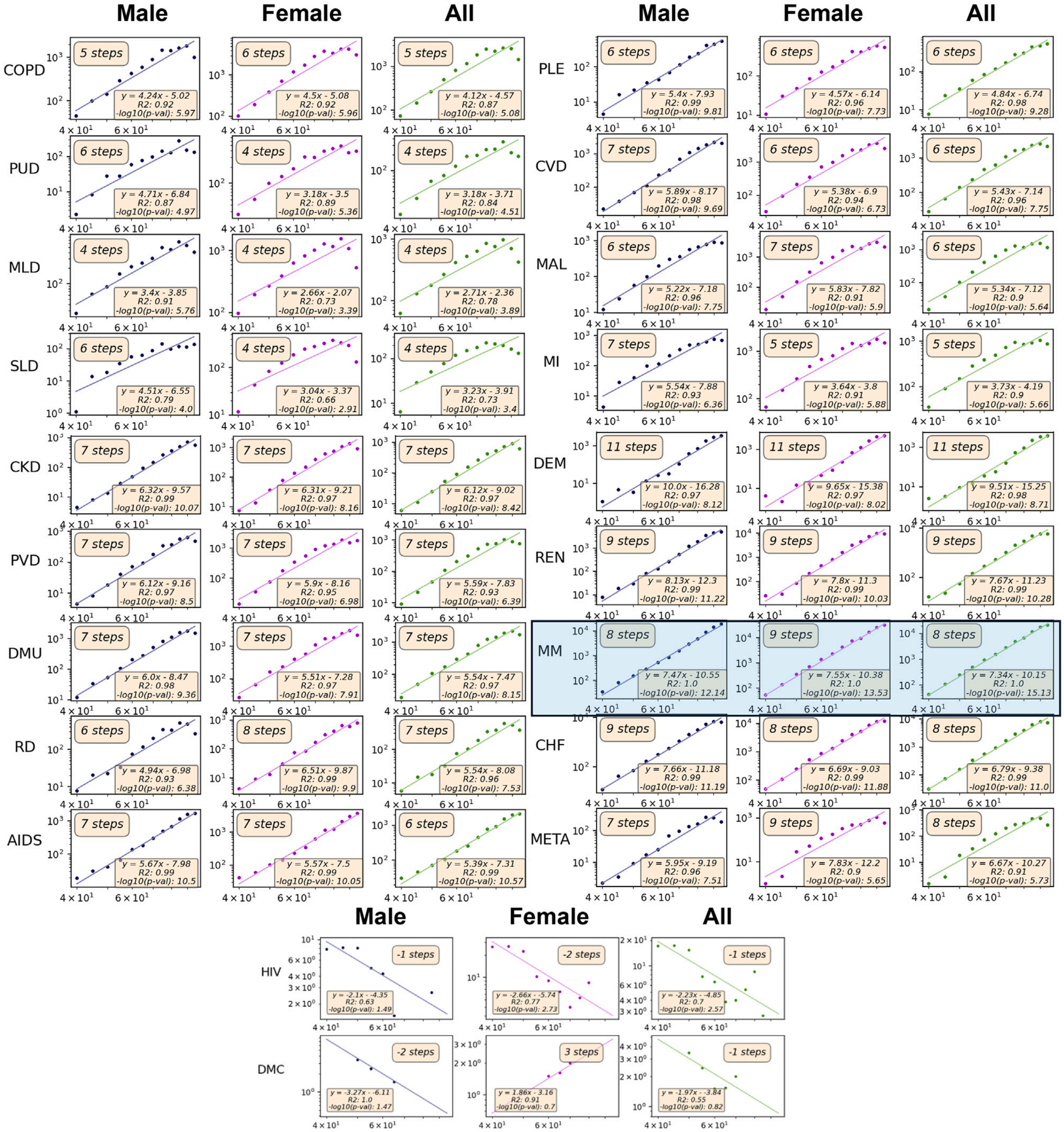


Figure 6. Multistep model analysis of CCI diseases age-stratified incidence profiles

Regression lines of the fit of the $\log_{10}$ of incidence (ordinates) vs. the $\log_{10}$ of age (abscissas). The continuous line is the regression line of the fit. Framed In boxes are framed the estimated number of steps and the fitted regression model, the R^2 of the regression, and the -$\log_{10}$(*p*-value) of the significance of the regression. The male, female, and pool of both sexes' cases are shown in blue, pink, and green color, respectively.

DISCUSSION

Most CCIs show the same trend of a slow increase in the number of cases until they reach a peak incidence between the ages of 60 and 85, depending on the disease, after which the incidence increases more rapidly. However, the dynamics of HIV are different from those of other diseases. The number of diagnoses increases rapidly until about the age of 35 and then it declines more slowly. This may be due to the infectious nature of the disease and its association with sexual practices that are less common in older age groups. It is noteworthy that the AIDS

Table 1. Main features of the multistep incidence-age models for the 19 CCI diseases and MM

D	Disease	CCI	Gender	Model of $\log_{10}$(incidence)	# steps	-$\log_{10}$(*p*-value)	R^2
DEM	Dementia	1	M	$10.0\cdot\log_{10}(\text{age})-16.28$	11	8.12	0.97
			F	$9.65\cdot\log_{10}(\text{age})-15.38$	11	8.02	0.97
			A	$9.51\cdot\log_{10}(\text{age})-15.25$	11	8.71	0.98
REN	Renal 1 Disease (Mild or Moderate)	1	M	$8.13\cdot\log_{10}(\text{age})-12.30$	9	11.22	0.99
			F	$7.80\cdot\log_{10}(\text{age})-11.30$	9	10.03	0.99
			A	$7.67\cdot\log_{10}(\text{age})-11.23$	9	10.28	0.99
MM	**Multimorbidity**	—	**M**	$\mathbf{7.47\cdot\log_{10}(\text{age})-10.55}$	**8**	**12.14**	**1.00**
			F	$\mathbf{7.55\cdot\log_{10}(\text{age})-10.38}$	**9**	**13.53**	**1.00**
			A	$\mathbf{7.34\cdot\log_{10}(\text{age})-10.15}$	**8**	**15.13**	**1.00**
CHF	Congestive Heart Failure	1	M	$7.62\cdot\log_{10}(\text{age})-11.11$	9	11.10	0.99
			F	$6.64\cdot\log_{10}(\text{age})-8.96$	8	11.63	0.99
			A	$6.75\cdot\log_{10}(\text{age})-9.32$	8	10.85	0.99
META	Metastatic Solid Tumor	6	M	$5.95\cdot\log_{10}(\text{age})-9.19$	7	7.51	0.96
			F	$7.83\cdot\log_{10}(\text{age})-12.20$	9	5.65	0.90
			A	$6.67\cdot\log_{10}(\text{age})-10.27$	8	5.73	0.91
CKD	Renal Disease (Severe)	3	M	$6.32\cdot\log_{10}(\text{age})-9.57$	7	10.07	0.99
			F	$6.31\cdot\log_{10}(\text{age})-9.21$	7	8.16	0.97
			A	$6.12\cdot\log_{10}(\text{age})-9.02$	7	8.42	0.97
PVD	Peripheral Vascular Disease	1	M	$6.12\cdot\log_{10}(\text{age})-9.16$	7	8.52	0.97
			F	$5.90\cdot\log_{10}(\text{age})-8.15$	7	6.99	0.95
			A	$5.59\cdot\log_{10}(\text{age})-7.83$	7	6.40	0.93
DMU	Diabetes without chronic complication	1	M	$6.0\cdot\log_{10}(\text{age})-8.47$	7	9.36	0.98
			F	$5.51\cdot\log_{10}(\text{age})-7.28$	7	7.91	0.97
			A	$5.54\cdot\log_{10}(\text{age})-7.47$	7	8.15	0.97
RD	Rheumatic Disease	1	M	$4.94\cdot\log_{10}(\text{age})-6.98$	6	6.38	0.93
			F	$6.51\cdot\log_{10}(\text{age})-9.87$	8	9.90	0.99
			A	$5.54\cdot\log_{10}(\text{age})-8.08$	7	7.53	0.96
AIDS	AIDS (HIV Infection + opportunistic infection)	6	M	$5.67\cdot\log_{10}(\text{age})-7.98$	7	10.50	0.99
			F	$5.57\cdot\log_{10}(\text{age})-7.50$	7	10.05	0.99
			A	$5.39\cdot\log_{10}(\text{age})-7.31$	6	10.57	0.99
PLE	Hemiplegia or Paraplegia	2	M	$5.40\cdot\log_{10}(\text{age})-7.93$	6	9.81	0.99
			F	$4.57\cdot\log_{10}(\text{age})-6.14$	6	7.73	0.96
			A	$4.84\cdot\log_{10}(\text{age})-6.74$	6	9.28	0.98
CVD	Cerebrovascular Disease	1	M	$5.89\cdot\log_{10}(\text{age})-8.17$	7	9.69	0.98
			F	$5.38\cdot\log_{10}(\text{age})-6.90$	6	6.73	0.94
			A	$5.43\cdot\log_{10}(\text{age})-7.14$	6	7.75	0.96
MAL	Any malignancy	2	M	$5.21\cdot\log_{10}(\text{age})-7.16$	6	7.77	0.96
			F	$5.79\cdot\log_{10}(\text{age})-7.74$	7	5.80	0.91
			A	$5.31\cdot\log_{10}(\text{age})-7.08$	6	5.60	0.90
MI	Myocardial Infarction	1	M	$5.54\cdot\log_{10}(\text{age})-7.88$	7	6.36	0.93
			F	$3.64\cdot\log_{10}(\text{age})-3.80$	5	5.88	0.91
			A	$3.73\cdot\log_{10}(\text{age})-4.19$	5	5.66	0.90
COPD	Chronic Pulmonary Disease	1	M	$4.24\cdot\log_{10}(\text{age})-5.02$	5	5.97	0.92
			F	$4.50\cdot\log_{10}(\text{age})-5.08$	6	5.96	0.92
			A	$4.12\cdot\log_{10}(\text{age})-4.57$	5	5.08	0.87

(*Continued on next page*)

Table 1. ***Continued***

D	Disease	CCI	Gender	Model of $\log_{10}$(incidence)	# steps	-$\log_{10}$(p-value)	R^2
PUD	Peptic Ulcer Disease	1	M	$4.71\cdot\log_{10}(\text{age})-6.84$	6	4.97	0.87
			F	$3.18\cdot\log_{10}(\text{age})-3.50$	4	5.36	0.89
			A	$3.18\cdot\log_{10}(\text{age})-3.71$	4	4.51	0.84
MLD	Mild Liver Disease	1	M	$3.40\cdot\log_{10}(\text{age})-3.85$	4	5.76	0.91
			F	$2.66\cdot\log_{10}(\text{age})-2.07$	4	3.39	0.73
			A	$2.71\cdot\log_{10}(\text{age})-2.36$	4	3.89	0.78
SLD	Moderate or Severe Liver Disease	3	M	$4.51\cdot\log_{10}(\text{age})-6.55$	6	4.00	0.79
			F	$3.04\cdot\log_{10}(\text{age})-3.37$	4	2.91	0.66
			A	$3.23\cdot\log_{10}(\text{age})-3.91$	4	3.40	0.73
HIV	HIV Infection, no AIDS	3	M	Not well-fitted model	−1	1.49	0.63
			F	Not well-fitted model	−2	2.73	0.77
			A	Not well-fitted model	−1	2.57	0.70
DMC	Diabetes with Chronic Complications	2	M	Not well-fitted model	−2	1.47	1.00
			F	$1.86\cdot\log_{10}(\text{age})-3.16$	3	0.70	0.91
			A	Not well-fitted model	−1	0.82	0.55

Diseases are ordered by descending number (#) of steps in the all cases combined data from both female and male subjects. D, disease acronym (also see Table S1); M, male; F, female; A, combined male and female.

case, which includes both the HIV infection and opportunistic infection, follows a well-fitted (-$\log_{10}$(p-value) > 10, R^2 = 0.99) seven-step multistage model for both sexes. This may indicate that the anergetic effect of the additional opportunistic infection on the HIV incidence remodels the data in a manner consistent with a multistep model.

We found that the MM, which includes all patients with comorbidity, fits very well (-$\log_{10}$(p-value) > 12, R^2 = 1) to a multistep model with eight steps for males, and nine steps for females. In fact, the MM data are better fitted by the multistep model than the diseases used to construct the CCI. Remarkably, this is an indication that the trigger of the individual CCI diseases includes several diseases in a synergistic manner to behave in a multistep manner, unlike the individual CCI diseases. Only DEM and REN, with 11 and 9 steps, respectively, have more steps than MM.

The disease with the highest number of steps is dementia (DEM), with 11 steps for both men and women. These results are consistent with the genealogy of NDs, as determined by a meta-analysis of age-stratified incidence data.[14,18] This analysis showed that several NDs form the crown of the ND tree, which includes NDs with 8 to 12 stages. We found that Metastatic Solid Tumor (META) in our dataset follows a multistep model with seven steps for males, nine for females, and eight for the pooled data of both sexes. This number of steps is higher than the five steps in other cancer models.[10] This discrepancy may be due to the fact that our dataset consists of patients with multimorbidity, with the patients with oncologic cancer in our dataset suffering from at least one additional disease. In addition, it could be due to the fact that we use ICD9 codes associated with metastasis (see Table S1), which correspond to a more advanced stage of cancer. However, malignancy (MAL), which corresponds to localized malignancies, follows a six-step multistep model in our study, which has only one step more than the general cancer case,[10] which could also be due to the fact that the MAL cases in our study suffer from at least one additional disease.

It is counterintuitive that diabetes without chronic complications (DMU) follows a well-fitted multistep model, while diabetes with chronic complications (DMC) does not follow a multistep model at all. It is possible that the relatively small number of DMC cases contributes to this apparent contradiction. The limited availability of DMC data may be due to the fact that only a small number of patients are diagnosed with both diabetes and a chronic complication at the same time. Typically, patients are first diagnosed with diabetes, and if another chronic complication develops, the physician adds additional conditions to the electronic medical record without necessarily classifying diabetes as DMC in those records. This case study highlights the challenges of accurately diagnosing diseases such as DMC, which require the detection of multiple elements that may not always be present simultaneously. Although we suspect that the DMC may follow a multistep model, the current method of recording DMC does not provide sufficient evidence to support this hypothesis.

From a gender perspective, we observed that in general in the CCI diseases, there were no significant differences in the distribution of the number of steps between the sexes in the different diseases (Table 1). Eight diseases have the same number of steps (DEM[1], REN[1], CKD[3], PVD[1], DMU[1], AIDS[6], PLE[2], MLD[1]), four have a difference of one step, two are higher in females, MAL[2] (6 steps for males, 7 for females), COPD[1] (5 steps for males, 6 for females) and two are higher in males: CHF[1] (9 steps for males, 8 for females), CVD[1] (7 steps for males, 6 for females). And five of the conditions have a difference of two steps, two higher in females: META[6] (7 steps for males, 9 for females), RD[1] (6 steps for males, 8 for females); and three higher in males: MI[1] (7 steps for males, 5 for females), PUD[1] (6 steps for males, 4 for females), and SLD[3] (6 steps for males, 4 for females). The remaining two diseases, HIV and DMC, do not follow a multistep model. Although females are better protected from infectious diseases[21] and more susceptible to autoimmune diseases,[22] in our study, at first

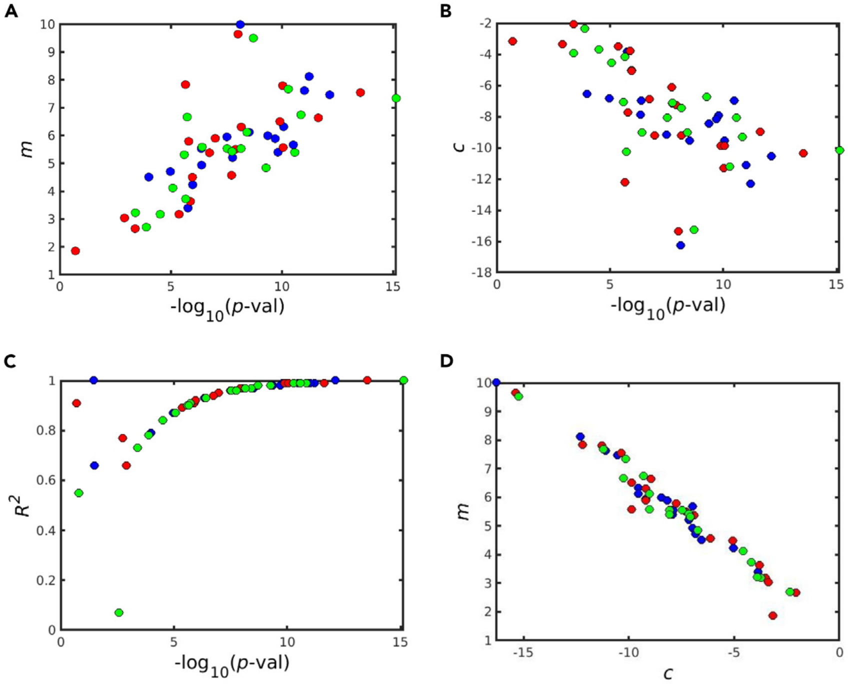


Figure 7. Scatterplots of the pairwise comparisons of fitness values and regression parameters of the multistep models of MM and CCI-associated diseases

(A) *m* (y axis) vs. the $\log_{10}$ of the *p*-value (x axis) of the regression fitness.

(B) *c*, intercept of the regression model vs. $\log_{10}$(*p*-value), $c = \log(u)$ where $u = \exp(c)$ is the background risk *u* of all steps.

(C) R^2 $\log_{10}$(*p*-value).

(D) *m* vs. *c*. The male, female, and pool of both sexes are shown in blue, red, and green color, respectively.

glance, there is no clear pattern in the gender distribution of the number of steps and the CCI score of the type of disease localization. However, in the case of COPD, comorbidities show significant differences by gender, namely chronic heart failure from the CCI diseases, edema, arterial hypertension, osteoporosis are more frequent in women, while ischemic heart disease from the CCI diseases, and alcoholism are more frequent in men[23] and in our study of COPD we found one step more in females than in males.

From an anatomical perspective, the central nervous system-related disease (DEM) has the highest number of steps (11), followed by kidney disease (REN, 9 steps; CKD, 7 steps), heart (CHF, 8 steps; CVD, 6 steps, MI 5 steps), peripheral vascular system (PVD, 7 steps), pancreas (DMU, 7 steps), joints (RD, 7 steps), cerebrovascular system (PLE, 6 steps; CVD 6 steps), lung (COPD, 5 steps), stomach (PUD, 4 steps), liver (MLD, 4 steps; SLD, 4 steps) (Figure 8).

The task of identifying an integrative explanatory model for this distribution of disease steps across anatomical systems represents a significant challenge. The number of steps appears to be homogeneous for diseases of the same organ, but heterogeneous for diseases of different organs. The number of steps varies considerably, from the highest number observed in the most complex organ of the body, which is also the most susceptible to failure with aging (the central nervous system), to the lowest number observed in the organ with the most regenerative capabilities, which is the most resilient to aging (the liver). The observed heterogeneity in the number of disease steps across different organs may also be related to the impact of lifestyle habits on the rate of organ deterioration. For example, in the lung, COPD is associated with smoking behavior, which has a cumulative effect across age, together with early life determinants that may play a role beyond smoking. The development of peptic ulcer disease (PUD) in the stomach is influenced by different dietary habits, while liver-related diseases are influenced by the varying levels of alcohol consumption throughout life.

In interpreting the multistep models, it is essential to recognize that each of the 19 CCI disease names serves as a proxy for a group of conditions, some of which are highly heterogeneous with different phenotypes. This is evident from the ICD column of Table S1. Furthermore, as previously discussed, in addition to these phenotypes, the disease progression may be influenced by environmental and lifestyle factors that evolve dynamically throughout life.

While Woolford et al.[24] assert that these patients can be regarded as a homogeneous population in terms of complexity, clinical vulnerability, frailty, mortality, functional impairment, polypharmacy, poor health-related quality of life, and a frequent situation of functional dependence, our study indicates that they are in fact heterogeneous in terms of the combinations of diseases present in each MM. Figure 2 illustrates the aforementioned heterogeneity. The figure illustrates the coexistence of at least two diseases, which collectively contribute to multimorbidity, forming a complex interaction network. The network encompasses diseases with varying numbers of steps, as well as diseases that do not adhere to a multistage model. It is noteworthy that the resulting MM ensemble is in close alignment with a multistage model

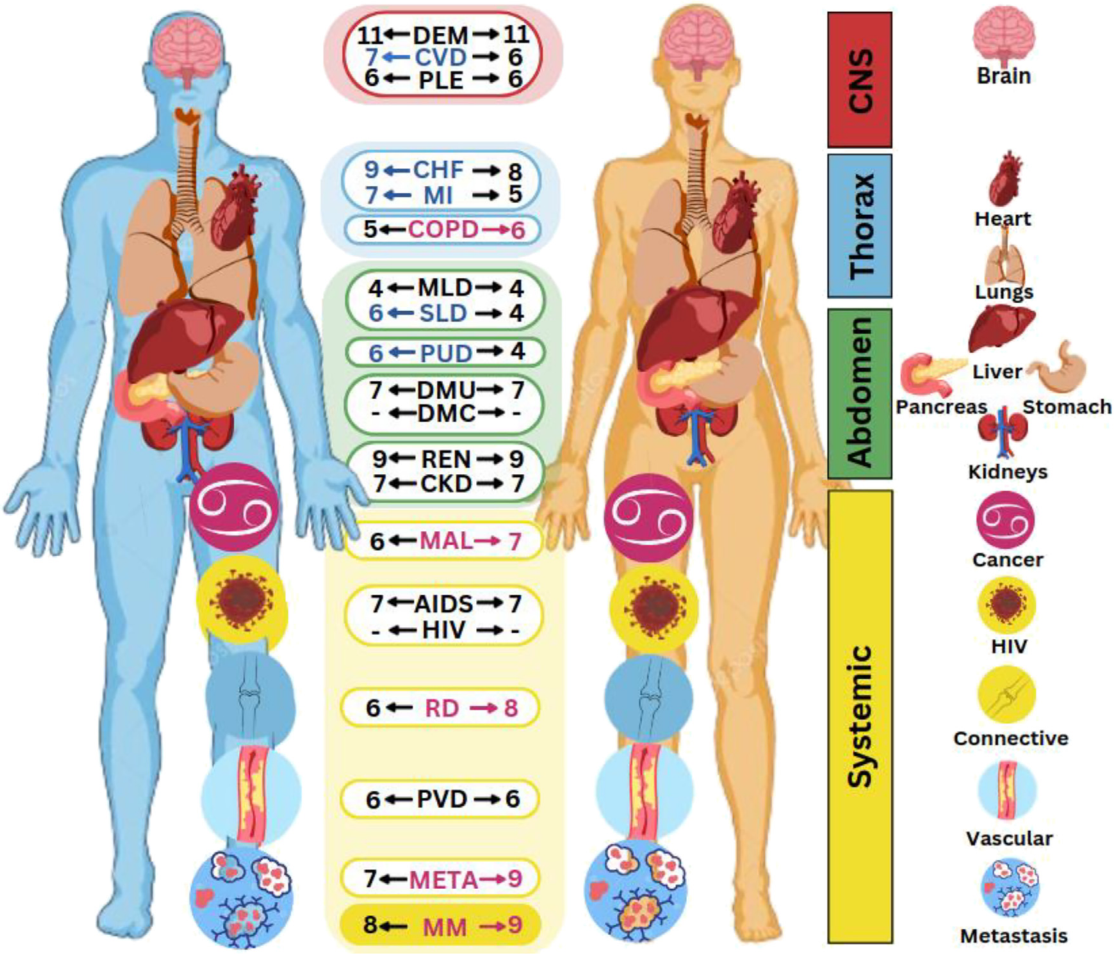


Figure 8. Distribution of the number of disease steps across the different parts of the body

Diseases that affect the same organ are framed together with a white background; the diseases that affect the same system are framed with the background of the system shown in the legend to the right of the figure. To the left and right of each disease are the number of steps for males and females, respectively. If the number of steps is higher for males than for females, the number of steps and the name of the disease are written in blue; if it is higher for females, it is written in pink.

comprising eight steps for the combined gender data. This model is consistent across both male and female subjects, with eight steps for males and nine for females.

Therefore, patients with multimorbidity display a high degree of heterogeneity. This is the reason why a number of researchers are engaged in the identification of multimorbidity patterns and the evolution or trajectories of these specific disease combinations. The objective is to identify more homogeneous subgroups that could be the target of specific interventions. The methodology we employed involved the decomposition of multimorbidity into its constituent disease components.

In the present work and in our previous study on the modeling of neurodegenerative diseases with multistep models,[18] we have observed that the diseases following a multistep model align approximately around an arc in the two-dimensional PCA representation. In contrast, diseases that do not adhere to such a model are not included in the aforementioned arc. Furthermore, we observed that in both cases, the diseases are approximately traversing the arc from higher to lower values of the 2nd principal component in conjunction with a reduction in the number of steps of each disease. It is yet to be determined whether this behavior is a general property associated with PCA of data with patterns or if it is specific to data that follow a multistep model.

These findings are observed at the macroscopic level, but further investigation may reveal a molecular-level correlation, which could be identified by the use of different technologies. These include studies that have identified evidence of disparate aging rates in different cell types or different tissues through bulk transcriptomics,[25] single-cell transcriptomics,[26–28] or longitudinal brain imaging and physiological phenotypes.[29]

Advantages and limitations of the study

The present study offers a comprehensive analysis of the incidence rates associated with chronic diseases and presents a stepwise model that explains the development of chronic diseases and multimorbidity. Such models provide an alternative perspective on multimorbidity that can be used in conjunction with scoring indices. The multistep view is more focused on the dynamics of the diseases, providing a detailed account of the number of steps required to trigger multimorbidity and the different diseases that contribute to it. This approach permits the categorization of the level of complexity—in terms of the number of steps required to trigger them—of the various diseases contributing to multimorbidity.

Once the number of steps has been determined, the subsequent step is to ascertain whether the triggering of the steps occurs in sequential order, using a Weibull model, or whether it follows a non-sequential model. Furthermore, it is essential to identify the specific steps involved. As with many diseases, the diseases associated with multimorbidity have genetic and environmental components that are combined in varying proportions. Consequently, some steps may have a genetic origin, while others may have an environmental etiology. Consequently, in order to ascertain which steps are requisite for the triggering of the diseases, it is essential to have access to genetic and environmental data. In order to ascertain the genetic-related steps, it is essential to have genetic data obtained from a range of costly genomic studies. Some progress has already been made in this direction.[30] In order to ascertain the environmental aspects, it is essential to have access to a substantial number of patients' clinical histories, lifestyles, and environmental data, which will enable the identification of the relevant steps with a sufficient degree of statistical significance.

It is crucial to consider the interplay between genetic and environmental factors. The influence of environmental conditions on the onset of disease may be contingent on the genetic predisposition. Even in cases where the disease is highly genetically associated, the final triggering event may be influenced by the number of environmental factors involved.

The aforementioned limitations are of the multistep models in general, other specific limitations of our present study are some diseases with low incidence, and thus with a small number of patients such as HIV or DMC may limit our modeling. However, we believe that it is appropriate to include such diseases in our study since they are used in the stratification method used by the Basque Health System.

Since the decrease in the population during aging depletes the last age ranges of patients, then in such ranges there are less data to fit the multistep regression model. We believe that since the linear behavior in the log scale of the data, the trends in lower, and middle ages could be extrapolated to higher ages, however, we must always keep in mind that the older ages of our models have less support of data.

Another limitation is that given our availability of clinical data our results and circumscribed to a small region of northern Spain with a relatively homogeneous population in relation to other regions, with prevalence rates of multimorbidity significantly lower than other regions. This limitation must be considered when trying to generalize our results. However, our results are important to establish a reference base for other studies in other regions with different levels of population heterogeneity.

Conclusions

A multistep model has been proposed for the first time to explain chronic multimorbidity. Indeed, the results showed that the onset of chronic multimorbidity can be effectively explained by a multistep model comprising eight steps for men, and nine steps for women, employing a systems biology approach. We deconvoluted the 19 comorbidities included in the calculation of the Charlson Comorbidity Index and found that six of them—namely congestive heart failure, AIDS, mild or moderate renal disease, hemiplegia or paraplegia, severe renal disease, and diabetes without chronic complications, used to construct the Charlson Comorbidity Index—also fit very well with multistep models with a number of steps ranging from nine to six.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to the lead contact, Prof. Marcos J. Araúzo-Bravo (mararabra@yahoo.co.uk).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The data post-processed for the 19 ICD diseases and for MM from the Basque Health System in semicolon column separated format have been deposited at GitLab and are publicly available as of the date of publication. Accession numbers are listed in the key resources table.
- The original source code in in Python version 2.7. to calculate the incidences of the patient data extratified by age for each of the diseases and to calculate the multistate models of each of the diseases have been deposited at GitLab and are publicly available as of the date of publication. Accession numbers are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this article is available from the lead contact upon request.

ACKNOWLEDGMENTS

We would like to thank Iván Araúzo-Gerovska for creating Figure 8 in Canva.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement N° 899417, by Ministerio de Ciencia e Innovación, Spain Grant No. PID2020-119715GB-I00/AEI/10.13039/501100011033, and by Instituto de Salud Carlos III, Infrastructure of Precision Medicine associated with Science and Technology (IMPaCT) of the Strategic Action in Health (iDATA-MP).

AUTHOR CONTRIBUTIONS

Conceptualization, M.J.A.-B.; methodology, M.A.E., D.G., and M.J.A.-B.; data curation, and M.A.E., M.S.-G., M.L.J.G., M.L.M.H., and M.J.A.-B.; investigation, M.A.E., D.G., M.S.-G., M.L.J.G., M.L.M.H., and M.J.A.-B.; formal analysis, M.A.E., D.G., M.S.-G., M.L.J.G., M.L.M.H., and M.J.A.-B.; writing – original draft, M.J.A.-B.; review and editing, M.A.E., D.G., M.S.-G., M.L.J.G., M.L.M.H., and M.J.A.-B.; visualization, M.A.E, D.G., and M.J.A.-B.; supervision, M.J.A.-B.; project administration, M.J.A.-B.; funding acquisition, M.J.A.-B. All authors have read, edited, and approved the final article.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Multimorbidity patient database construction
  - Multimorbidity (MM) and Charlson Comorbidity Index (CCI) diseases
  - Global analysis of disease incidence versus age data
  - Algorithm for calculating multistep models of disease incidence versus age data
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.110816.

Received: April 7, 2024
Revised: July 13, 2024
Accepted: August 22, 2024
Published: August 26, 2024

REFERENCES

1. Marengoni, A., Angleman, S., Melis, R., Mangialasche, F., Karp, A., Garmen, A., Meinow, B., and Fratiglioni, L. (2011). Aging with multimorbidity: a systematic review of the literature. Ageing Res. Rev. *10*, 430–439. https://doi.org/10.1016/j.arr.2011.03.003.
2. Barnett, K., Mercer, S.W., Norbury, M., Watt, G., Wyke, S., and Guthrie, B. (2012). Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. Lancet *380*, 37–43. https://doi.org/10.1016/S0140-6736(12)60240-2.
3. Ho, I.S., Azcoaga-Lorenzo, A., Akbari, A., Davies, J., Hodgins, P., Khunti, K., Kadam, U., Lyons, R., McCowan, C., Mercer, S.W., et al. (2022). Variation in the estimated prevalence of multimorbidity: systematic review and meta-analysis of 193 international studies. BMJ Open *12*, e057017. https://doi.org/10.1136/bmjopen-2021-057017.
4. McPhail, S.M. (2016). Multimorbidity in chronic disease: impact on health care resources and costs. Risk Manag. Healthc. Policy *9*, 143–156. https://doi.org/10.2147/RMHP.S97248.
5. Skou, S.T., Mair, F.S., Fortin, M., Guthrie, B., Nunes, B.P., Miranda, J.J., Boyd, C.M., Pati, S., Mtenga, S., and Smith, S.M. (2022). Multimorbidity. Nat. Rev. Dis. Primers *8*, 48. https://doi.org/10.1038/s41572-022-00376-4.
6. de Groot, V., Beckerman, H., Lankhorst, G.J., and Bouter, L.M. (2003). How to measure comorbidity: a critical review of available methods. J. Clin. Epidemiol. *56*, 221–229. https://doi.org/10.1016/S0895-4356(02)00585-1.
7. Elixhauser, A., Steiner, C., Harris, D.R., and Coffey, R.M. (1998). Comorbidity measures for use with administrative data. Med. Care *36*, 8–27. https://doi.org/10.1097/00005650-199801000-00004.
8. Alvarez-Galvez, J., and Vegas-Lozano, E. (2022). Discovery and classification of complex multimorbidity patterns: unravelling chronicity networks and their social profiles. Sci. Rep. *12*, 20004. https://doi.org/10.1038/s41598-022-23617-8.
9. Fisher, J.C., and Hollomon, J.H. (1951). A hypothesis for the origin of cancer foci. Cancer *4*, 916–918. https://doi.org/10.1002/1097-0142(195109)4:5<916::AID-CNCR2820040504>3.0.CO;2-7.
10. Armitage, P., and Doll, R. (1954). The age distribution of cancer and a multi-stage theory of carcinogenesis. Br. J. Cancer *8*, 1–12. https://doi.org/10.1038/bjc.1954.1.
11. Al-Chalabi, A., Calvo, A., Chio, A., Colville, S., Ellis, C.M., Hardiman, O., Heverin, M., Howard, R.S., Huisman, M.H.B., Keren, N., et al. (2014). Analysis of amyotrophic lateral sclerosis as a multistep process: a population-based modelling study. Lancet Neurol. *13*, 1108–1113. https://doi.org/10.1016/S1474-4422(14)70219-4.
12. Chiò, A., Mazzini, L., D'Alfonso, S., Corrado, L., Canosa, A., Moglia, C., Manera, U., Bersano, E., Brunetti, M., Barberis, M., et al. (2018). The multistep hypothesis of ALS revisited: the role of genetic mutations. Neurology *91*, e635–e642. https://doi.org/10.1212/WNL.0000000000005996.
13. Licher, S., van der Willik, K.D., Vinke, E.J., Yilmaz, P., Fani, L., Schagen, S.B., Ikram, M.A., and Ikram, M.K. (2019). Alzheimer's disease as a multistage process: an analysis from a population-based cohort study. Aging *11*, 1163–1176. https://doi.org/10.18632/aging.101816.
14. Gerovska, D., Irizar, H., Otaegi, D., Ferrer, I., López de Munain, A., and Araúzo-Bravo, M.J. (2020). Genealogy of the neurodegenerative diseases based on a meta-analysis of age-stratified incidence data. Sci. Rep. *10*, 18923. https://doi.org/10.1038/s41598-020-75014-8.
15. Garton, F.C., Trabjerg, B.B., Wray, N.R., and Agerbo, E. (2021). Cardiovascular disease, psychiatric diagnosis and sex differences in the multistep hypothesis of amyotrophic lateral sclerosis. Eur. J. Neurol. *28*, 421–429. https://doi.org/10.1111/ene.14554.
16. Vucic, S., Higashihara, M., Sobue, G., Atsuta, N., Doi, Y., Kuwabara, S., Kim, S.H., Kim, I., Oh, K.W., Park, J., et al. (2020). ALS is a multistep process in South Korean, Japanese, and Australian patients. Neurology *94*, e1657–e1663. https://doi.org/10.1212/WNL.0000000000009015.
17. Le Heron, C., MacAskill, M., Mason, D., Dalrymple-Alford, J., Anderson, T., Pitcher, T., and Myall, D. (2021). A multi-step model of Parkinson's disease pathogenesis. Mov. Disord. *36*, 2530–2538. https://doi.org/10.1002/mds.28719.
18. Gerovska, D., and Araúzo-Bravo, M.J. (2022). The common incidence-age multistep model of neurodegenerative diseases revisited: wider general age range of incidence corresponds to fewer disease steps. Cell Biosci. *12*, 11. https://doi.org/10.1186/s13578-021-00737-8.
19. Webster, J.A., and Clarke, R. (2022). Sporadic, late-onset, and multistage diseases. PNAS Nexus *1*, pgac095. https://doi.org/10.1093/pnasnexus/pgac095.

20. Kalbfleisch, J.D., and Schaubel, D.E. (2023). Fifty Years of the Cox Model. Annual Review of Statistics and Its Application *10*, 1–23. https://doi.org/10.1146/annurev-statistics-033021-014043.
21. van, L., and Altfeld, M. (2014). Sex differences in infectious diseases-common but neglected. J. Infect. Dis. *209*, S79–S80. https://doi.org/10.1093/infdis/jiu159.
22. Migliore, L., Nicolì, V., and Stoccoro, A. (2021). Gender specific differences in disease susceptibility: The role of epigenetics. Biomedicines *9*, 652. https://doi.org/10.3390/biomedicines9060652.
23. Almagro, P., López García, F., Cabrera, F.J., Montero, L., Morchón, D., Díez, J., and Soriano, J.B.; Grupo Epoc De La Sociedad Española De Medicina Interna (2010). Comorbidity and gender-related differences in patients hospitalized for COPD. The ECCO study. Respir. Med. *104*, 253–259. https://doi.org/10.1016/j.rmed.2009.09.019.
24. Woolford, S.J., Aggarwal, P., Sheikh, C.J., and Patel, H.P. (2021). Frailty, multimorbidity and polypharmacy. Medicine *49*, 166–172. https://doi.org/10.1016/j.mpmed.2020.12.010.
25. Schaum, N., Lehallier, B., Hahn, O., Pálovics, R., Hosseinzadeh, S., Lee, S.E., Sit, R., Lee, D.P., Losada, P.M., Zardeneta, M.E., et al. (2020). Ageing hallmarks exhibit organ-specific temporal signatures. Nature *583*, 596–602. https://doi.org/10.1038/s41586-020-2499-y.
26. Lagger, C., and de Magalhães, J.P. (2022). Single-cell gene regulation across aging tissues. Nat. Aging *2*, 468–470. https://doi.org/10.1038/s43587-022-00238-4.
27. Tabula Muris Consortium (2020). A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. Nature *583*, 590–595. https://doi.org/10.1038/s41586-020-2496-1.
28. Ibañez-Solé, O., Ascensión, A.M., Araúzo-Bravo, M.J., and Izeta, A. (2022). Lack of evidence for increased transcriptional noise in aged tissues. Elife *11*, e80380. https://doi.org/10.7554/eLife.80380.
29. Tian, Y.E., Cropley, V., Maier, A.B., Lautenschlager, N.T., Breakspear, M., and Zalesky, A. (2023). Heterogeneous aging across multiple organ systems and prediction of chronic disease and mortality. Nat. Med. *29*, 1221–1231. https://doi.org/10.1038/s41591-023-02296-6.
30. Dong, G., Feng, J., Sun, F., Chen, J., and Zhao, X.M. (2021). A global overview of genetically interpretable multimorbidities among common diseases in the UK Biobank. Genome Med. *13*, 110.
31. Soto-Gordoa, M., de Manuel, E., Fullaondo, A., Merino, M., Arrospide, A., Igartua, J.I., Mar, J., and CareWell, G. (2019). Impact of stratification on the effectiveness of a comprehensive patient-centered strategy for multimorbid patients. Health Serv. Res. *54*, 466–473. https://doi.org/10.1111/1475-6773.13094.
32. Sicras-Mainar, A., Serrat-Tarrés, J., Navarro-Artieda, R., Llausí-Sellés, R., Ruano-Ruano, I., and González-Ares, J.A. (2007). Adjusted Clinical Groups use as a measure of the referrals efficiency from primary care to specialized in Spain. Eur. J. Publ. Health *17*, 657–663. https://doi.org/10.1093/eurpub/ckm044.
33. Glasheen, W.P., Cordier, T., Gumpina, R., Haugh, G., Davis, J., and Renda, A. (2019). Charlson Comorbidity Index: *ICD-9* Update and *ICD-10* Translation. Am. Health Drug Benefits *12*, 188–197.
34. Charlson, M.E., Pompei, P., Ales, K.L., and MacKenzie, C.R. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J. Chron. Dis. *40*, 373–383. https://doi.org/10.1016/0021-9681(87)90171-8.
35. Charlson, M.E., Carrozzino, D., Guidi, J., and Patierno, C. (2022). Charlson Comorbidity Index: A critical review of clinimetric properties. Psychother. Psychosom. *91*, 8–35. https://doi.org/10.1159/000521288.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
AIDS.csv (Acquired Immunodeficiency Syndrome, HIV Infection + opportunistic infection) patient data post-processed from the Basque Health System (BHS), in semicolon column separated format.	This paper	https://gitlab.com/mikel_arrospide/multistep_model/-/blob/main/input/AIDS.csv?ref_type=heads
CHF.csv (Congestive Heart Failure) patient data post-processed from the Basque Health System (BHS), in semicolon column separated format.	This paper	https://gitlab.com/mikel_arrospide/multistep_model/-/blob/main/input/CHF.csv?ref_type=heads
CKD.csv (Renal Disease, Severe) patient data post-processed from the Basque Health System (BHS), in semicolon column separated format.	This paper	https://gitlab.com/mikel_arrospide/multistep_model/-/blob/main/input/CKD.csv?ref_type=heads
COPD.csv (Chronic Pulmonary Disease) patient data post-processed from the Basque Health System (BHS), in semicolon column separated format.	This paper	https://gitlab.com/mikel_arrospide/multistep_model/-/blob/main/input/COPD.csv?ref_type=heads
CVD.csv (Cerebrovascular Disease) patient data post-processed from the Basque Health System (BHS), in semicolon column separated format.	This paper	https://gitlab.com/mikel_arrospide/multistep_model/-/blob/main/input/CVD.csv?ref_type=heads
DEM.csv (Dementia) patient data post-processed from the Basque Health System (BHS), in semicolon column separated format.	This paper	https://gitlab.com/mikel_arrospide/multistep_model/-/blob/main/input/DEM.csv?ref_type=heads
DMC.csv (Diabetes with Chronic Complications) patient data post-processed from the Basque Health System (BHS), in semicolon column separated format.	This paper	https://gitlab.com/mikel_arrospide/multistep_model/-/blob/main/input/DMC.csv?ref_type=heads
DMU.csv (Diabetes without Chronic Complications) patient data post-processed from the Basque Health System (BHS), in semicolon column separated format.	This paper	https://gitlab.com/mikel_arrospide/multistep_model/-/blob/main/input/DMU.csv?ref_type=heads
HIV.csv (Human Immunodeficiency Virus infection, no AIDS) patient data post-processed from the Basque Health System (BHS), in semicolon column separated format.	This paper	https://gitlab.com/mikel_arrospide/multistep_model/-/blob/main/input/HIV.csv?ref_type=heads
MAL.csv (Any malignancy) patient data post-processed from the Basque Health System (BHS), in semicolon column separated format.	This paper	https://gitlab.com/mikel_arrospide/multistep_model/-/blob/main/input/MAL.csv?ref_type=heads
META.csv (Metastatic Solid Tumor) patient data post-processed from the Basque Health System (BHS), in semicolon column separated format.	This paper	https://gitlab.com/mikel_arrospide/multistep_model/-/blob/main/input/META.csv?ref_type=heads
MI.csv (Myocardial Infarction) patient data post-processed from the Basque Health System (BHS), in semicolon column separated format.	This paper	https://gitlab.com/mikel_arrospide/multistep_model/-/blob/main/input/MI.csv?ref_type=heads
MLD.csv (Mild Liver Disease) patient data post-processed from the Basque Health System (BHS), in semicolon column separated format.	This paper	https://gitlab.com/mikel_arrospide/multistep_model/-/blob/main/input/MLD.csv?ref_type=heads
MM.csv (Multimorbidity) patient data post-processed from the Basque Health System (BHS), in semicolon column separated format.	This paper	https://gitlab.com/mikel_arrospide/multistep_model/-/blob/main/input/MM.csv?ref_type=heads

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
PLE.csv (Hemiplegia or Paraplegia) patient data post-processed from the Basque Health System (BHS), in semicolon column separated format.	This paper	https://gitlab.com/mikel_arrospide/multistep_model/-/blob/main/input/PLE.csv?ref_type=heads
PUD.csv (Peptic Ulcer Disease) patient data post-processed from the Basque Health System (BHS), in semicolon column separated format.	This paper	https://gitlab.com/mikel_arrospide/multistep_model/-/blob/main/input/PUD.csv?ref_type=heads
PVD.csv (Peripheral Vascular Disease) patient data post-processed from the Basque Health System (BHS), in semicolon column separated format.	This paper	https://gitlab.com/mikel_arrospide/multistep_model/-/blob/main/input/PVD.csv?ref_type=heads
RD.csv (Rheumatic Disease) patient data post-processed from the Basque Health System (BHS), in semicolon column separated format.	This paper	https://gitlab.com/mikel_arrospide/multistep_model/-/blob/main/input/RD.csv?ref_type=heads
REN.csv (Renal 1 Disease, Mild or Moderate) patient data post-processed from the Basque Health System (BHS), in semicolon column separated format.	This paper	https://gitlab.com/mikel_arrospide/multistep_model/-/blob/main/input/REN.csv?ref_type=heads
SLD.csv (Moderate or Severe Liver Disease) patient data post-processed from the Basque Health System (BHS), in semicolon column separated format.	This paper	https://gitlab.com/mikel_arrospide/multistep_model/-/blob/main/input/SLD.csv?ref_type=heads
Software and Algorithms		
Get_Incidences.py Software code in Python version 2.7. To get the patient data stratified by age for each of the diseases.	This paper	https://gitlab.com/mikel_arrospide/multistep_model/-/blob/main/Get_Incidences.py?ref_type=heads
MultiStep_model.py Software code in Python version 2.7. To calculate the multistate models of each of the diseases.	This paper	https://gitlab.com/mikel_arrospide/multistep_model/-/blob/main/MultiStep_model.py?ref_type=heads

EXPERIMENTAL MODEL AND SUBJECT DETAILS

We used a retrospective approach and collected data on patients classified as having chronic multimorbidity from the Oracle Business Intelligence (OBI) database of the Basque Health System (BHS), Spain, between January 1, 2014, and March 31, 2021.

METHOD DETAILS

Multimorbidity patient database construction

Patient stratification

The BHS classified patients as MM following a risk stratification for case finding using a strategy launched in 2011. The primary goal of this initiative was to prioritize complicated, fragile, and high-risk patients based on Adjusted Clinical Groups (ACG). Patient prioritization was achieved through the implementation a specific care pathway.[31] The strategy relied on health administrative data from both primary care and hospitals, as well as a list of 52 chronic conditions. This approach has resulted in a comprehensive dataset that integrates information from a variety of sources, including primary and specialty care records, census data, and clinical data. The ACG method[32] was used to categorize individuals based on their disease burden.

The multimorbidity stratification system used by BHS classifies the population based on the probability for each person of consuming health resources in the next year. To do this, a Predictive Index is calculated for each person. This is an index that, considering a person's consumption of resources and services in the previous year, predicts the level of resource consumption next year. The three strata, in which people with chronic illness are classified, are determined based on the Predictive Index of each person with a strategic purpose: (i) Patients in the red stratum: they are characterized as having a high probability of resource consumption (admissions, care consultations in the emergency room or in primary care consultations). (ii) Patients in the orange stratum: they are characterized as having a medium probability of resource consumption. (iii) Patients in the yellow stratum: they are characterized as a having low probability of resource consumption. This process of patient identification is objective and can be transferred to other settings with universalized national systems. As MM patients we took the patients from the three strata.

Patient filtering

The total population of the region during the period under study was 2,172,877. The stratification method identified a total of 42,850 patients with MM. (i) we discarded 9,745 patients classified as MM before 2014 since the classification method was not consistent with the current one due to changes in the design of the predictive index used to stratify the patients. (ii) we excluded further 1,383 cases of postmortem MM, as in 2014 the BHS classified patients as MM even when the patients were deceased, provided they met the MM criteria. The BHS stratification method does not require a minimum number of days for a patient to be classified as MM. However, given the chronicity of multiple diseases used in the system and the condition that a patient must suffer from at least two diseases, we expect that the proportion of sporadic cases with very few days in the dataset will be very low, as we have discarded in this stage the postmortem cases. (iii) for calculation of the incidence rates, we implemented two strategies: (a) For the MM global case, we excluded patients classified as having MM who were younger than 35 years or older than 95 years. We used this limit in accordance with the recommendation of the expert curators of the BHS database. Accordingly, we grouped patients by age based on five-year ranges (40–45], (45–50], …, (85–90], (90, 95], i.e., with 38-year-olds falling into the 40-year-old group and 37-year-olds in the 35-year-old group. (b) With regard to the 19 diseases associated with the Charlson Comorbidity Index (CCI) as presented in Table S1, it was deemed necessary to relax the age limit for diseases with low prevalence in order to increase the population of such diseases. Following the integration of the two strategies, a reduction in the number of patients was observed, from 31,722 to 31,415. This represents a reduction of less than 1% (307 patients) in the final filtering stage (see Figure 1).

To adjust for demographic values, we used data from the Basque Institute of Statistics (Eustat) on the general population of the Basque Country. We extracted information on age, sex, diseases, and age of diagnosis from the OBI database for each patient.

Multimorbidity (MM) and Charlson Comorbidity Index (CCI) diseases

To investigate a possible multistep model incidence-age relationship of the MM, we did not focus solely on BHS diagnosed as MM by the ACG method. Instead, to disentangle the MM diseases components, but keeping diseases with enough number of patients, form the original list of 52 diseases we collected the 19 diseases used to calculate the scoring scheme CDMF (Claims-based, Disease-specific refinements, Matching translation to ICD10, Flexibility)[33] for the new the Charlson Comorbidity Index (CCI).[34,35] The 19 diseases are summarized in Table S1: {Myocardial Infarction (MI), Congestive Heart Failure (CHF), Peripheral Vascular Disease (PVD), Cerebrovascular Disease (CVD), Dementia (DEM), Chronic Pulmonary Disease (COPD), Rheumatic Disease (RD), Peptic Ulcer Disease (PUD), Mild Liver Disease (MLD), Diabetes without chronic complication (DMU), Renal l Disease (Mild or Moderate) (REN), Diabetes with Chronic Complications (DMC), Hemiplegia or Paraplegia (PLE), Any malignancy (MAL), Moderate or Severe Liver Disease (SLD), Renal Disease (Severe) (CKD), HIV Infection, no AIDS (HIV), Metastatic Solid Tumor (META), AIDS (HIV Infection + opportunistic infection) (AIDS)}. Each disease is assigned a value of 1, 2, 3, or 6, based on the increasing risk of mortality associated with it. For simplicity, we refer to this list of 19 diseases from here on, as a list of CCI diseases.

The incidence of each of the 20 pathologies, including the MM and the 19 CCI-associated diseases, was calculated per 100,000 inhabitants using Eustat demographic data (https://www.eustat.eus/bankupx/pxweb/es/DB/-/PX_010154_cepv1_ep10b.px/table/tableViewLayout1/).

Global analysis of disease incidence versus age data

In order to obtain a comprehensive overview of all the diseases under analysis, we performed a global analysis. To carry up such an analysis, it is necessary to align the incidence data for all diseases in order to ensure the same range of age incidence measurements, resulting in a matrix that included all data for the same age ranges. We binned the ages of all analyzed patients from 40 to 95 in 5-year intervals and calculated the disease incidence per 100,000 population for each age group. The resulting matrix was transformed using the *z*-sores of each raw and subjected to principal component analysis (PCA) for multidimensional reduction. The *z*-scores were calculated for each row by subtracting the row mean from the value and then dividing the result by the row standard deviation. Hierarchical clustering with the standardized Euclidean metric and the Ward linkage method were used to demonstrate potential relationships between the dynamics of disease incidence dynamics. To calculate the standardized Euclidean metric, each coordinate difference between observations is scaled by dividing by the corresponding element of the standard deviation metric.

Algorithm for calculating multistep models of disease incidence versus age data

Multistep regression models to calculate the number of steps for MM and each CCI disease

We fitted a regression equation for each disease incidence against the age profile of each dataset. To determine whether the pathogenesis of a disease followed a multistep model, we used a logarithmic transformation of the regression equation, as in Armitage and Doll.[10] This transformation was chosen because age and incidence on a logarithmic scale must fit a linear regression, and the slope of the regression is directly related to the average number of steps, which is calculated as slope +1. The incidence rate is the number of new cases per population at risk during a given time period. For a multistep model, the incidence (i) over time (t) is calculated using the following formula: $i = u_1 \cdot u_2 \cdot u_3 \cdot \ldots u_{n-1} \cdot u_n \cdot t(n-1)$, where u_k is the average background risk of step k. The regression line in logarithmic scale of i across t is $\log(i) = (n-1) \cdot \log_{10}(t) + c$, where $n - 1 = m$ is the slope of the regression line, $n = m + 1$ is the number of steps, and $c = \log_{10}(u_1 \cdot u_2 \cdot u_3 \cdot \ldots u_{n-1} \cdot u_n) = \log_{10}(u)$ is the intercept of the regression line. The background risk, u, of all steps is determined by taking the exponential of the intercept $u = \exp(c)$, where $c = \log_{10}(u_1 \cdot u_2 \cdot u_3 . \ldots u_{n-1} \cdot u_n) = \log_{10}(u)$. The geometric mean background risk for of all steps is $\mu(u) = u/n$.

In these models the predictive variables are the vector $\log_{10}$(age) and the respond variables the vector of the vector $\log_{10}$(incidence). We build the models for each of the 19 CCI diseases independently. For the MM we build global models for all the data. For each of the analyze 20 disease we build three independent models for female, male and the combined case of both genders.

QUANTIFICATION AND STATISTICAL ANALYSIS

To determine the statistical significance of the linear regression relationship between the response variable and the predictor variables, we used the R^2 coefficient of determination and the *p*-value for the F-test on the regression model. A *p*-value of less than 0.05 is considered significant, indicating that the multistep model is valid. These values are represented in Figure 7 and listed in Table 1.