WIREs Wiley Interdisciplinary Reviews

A Revision and Analysis of the Comprehensiveness of the Main Longitudinal Studies of Human Ageing for Data Mining Research

Article Type:

○ OPINION ○ OVERVIEW ○ FOCUS ARTICLE

○ PRIMER ◉ ADVANCED REVIEW ○ SOFTWARE FOCUS

Authors:

First author Caio Eduardo Ribeiro; Pontifícia Universidade Católica de Minas Gerais; caioedurib@gmail.com
Second author Luis Henrique S. Brito; Pontifícia Universidade Católica de Minas Gerais
Third author Cristiane Neri Nobre; Pontifícia Universidade Católica de Minas Gerais
Fourth author Alex A. Freitas; University of Kent
Fifth author Luis Enrique Zárate; Pontifícia Universidade Católica de Minas Gerais

Abstract

Human ageing is a global problem that will have a large socio-economic impact. A better understanding of ageing can direct public policies that minimize its negative effects in the future. Over many years, several longitudinal studies of human ageing have been conducted aiming to comprehend the phenomenon, and various factors influencing human ageing are under analysis. In this review, we categorize the main aspects affecting human ageing into a taxonomy for assisting data mining research on this topic. We also present tables summarizing the main characteristics of 64 research articles using data from ageing-related longitudinal studies, in terms of the ageing-related aspects analysed, the main data analysis techniques used, and the specific longitudinal database mined in each article. Finally, we analyse the comprehensiveness of the main databases of longitudinal studies of human ageing worldwide, regarding which proportion of the proposed taxonomy's aspects are covered by each longitudinal database. We observed that most articles analysing such data are using classical (parametric, linear) statistical techniques, with little use of more modern (non-parametric, non-linear) data mining methods for analysing longitudinal databases of human ageing. We hope that this article will contribute to data mining research in two ways: first, by drawing attention to the important problem of global ageing and the free availability of several longitudinal databases of human ageing; second, by providing useful information to make research design choices about mining such data, e.g. which longitudinal study and which types of ageing-related aspects should be analysed, depending on the research's goals.

1 - Introduction

A longitudinal study aims to observe the changes on variables related to a problem domain, to the same sample objects, through previously established periods of time, named waves[15]. This kind of study tries to find cause and effect relations between the different values assigned to variables, considering the temporal aspect.

The longitudinal data analysis techniques most frequently used in the literature are based on classical statistics, such as hypothesis tests, correlation analysis, linear regression and logistic regression analysis. Note that such classical statistical data analysis techniques are usually parametric (making strong assumptions about the data distribution) and often detect only linear correlations in data. By contrast, the use of more flexible non-parametric, non-linear data mining or machine learning techniques – e.g., support vector machines – have been much less explored in the context of longitudinal data, although such techniques have also been adapted to this type of data – e.g. Du[16], and Chen[12].

In the context of longitudinal studies, an area that has been gaining attention from the scientific community and governmental agencies are the human ageing studies. It is estimated that the elderly population will surpass 21.5% of the global population by 2050, which will strongly impact society and economy[68]. To understand this impact, research projects from different areas have been investigating a large number of different aspects of the ageing process.

In this context, as one of the contributions of this article, the main aspects associated with human ageing in the literature were identified. These aspects have been organized into a taxonomy for facilitating the analysis of data contained in the main longitudinal studies on human ageing. This taxonomy is also used to study how comprehensive each longitudinal study is, in terms of the proportion of types of aspects covered by the study. As a second contribution, we investigated which data analysis techniques are most frequently applied to longitudinal studies.

This review article aims to support future Longitudinal Data Mining (LDM) research projects on human ageing which use data sources such as the English Longitudinal Study of Ageing (ELSA), the Survey of Health Ageing and Retirement in Europe (SHARE), the Chinese Longitudinal Healthy Longevity Survey (CLHLS), etc. To the best of our knowledge, there is just one other systematic review on longitudinal studies on human ageing, namely the review by Kaiser[31], which is conceptually quite different from our current review. More precisely, Kaiser's review is more focused on the ageing background and the methodology used to create many longitudinal databases of human ageing, with not much emphasis on data mining. By contrast, our review is more focused on the issue of how such longitudinal databases of ageing can be used for data mining and knowledge discovery purposes.

This review article is organized as follows: Section 2 presents general background on human ageing and knowledge discovery in longitudinal databases of ageing. Sections 3 and 4 present, respectively, the proposed taxonomy of many aspects affecting ageing (based on an extensive literature review) and the previously mentioned analysis of the degree of comprehensiveness of each of the main longitudinal studies of ageing with respect to the aspects identified in that taxonomy. Section 5 discusses the use of data analysis techniques on longitudinal databases of ageing. Finally, Section 6 concludes the article with our considerations regarding the perspectives and challenges of longitudinal data mining applied on human ageing studies.

2 – Background on Ageing and Knowledge Discovery in Longitudinal Databases

2.1 – Human Ageing at the Population Level

The global ageing phenomenon is caused by an increased life expectancy allied with the decline of birth rates. These facts have been occurring throughout countries all over the world, and trigger a gradual increase of the proportion of elderly citizens in relation to the rest of the population. According to data from the World Population Prospects: The 2015 Revision[68], the proportion of the population 60 or more years old, which currently represents 12.3% of the global population, is estimated to reach 21.5% by 2050, and 28.3% by 2100 (Figure 1).

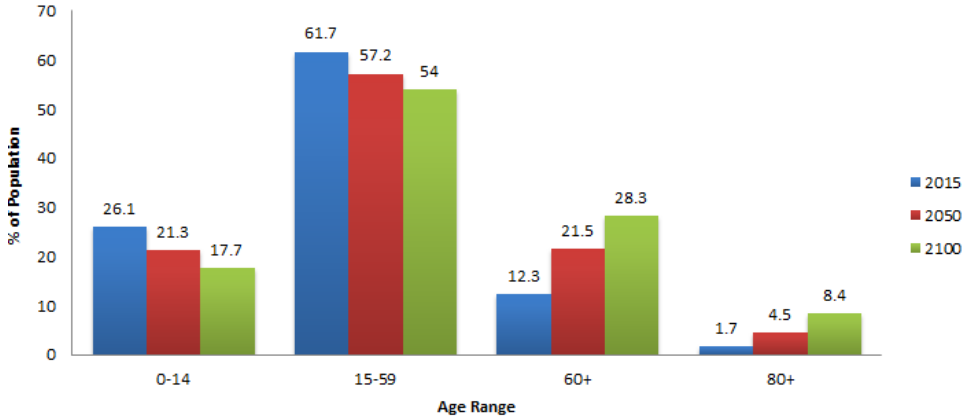

Figure 1. Proportion of population by age range. Source: UNDESA[68]

Populational ageing impacts the whole structure of our society, especially regarding social security issues, because the ratio of active workers over retired workers will suffer a decline, with several socioeconomic implications[43]. Therefore, understanding the human ageing process is of interest to society as a whole, because it can help guiding the creation of public policies aimed at the older segment of the society, besides helping fighting the cognitive losses and many age-related diseases. The human ageing study is highly interdisciplinary, gathering efforts from different areas of knowledge[6], such as biology, medicine, psychology, social sciences, economy, etc. Accordingly, in this review paper we identify a wide variety of factors affecting ageing, from physical and mental health factors to psychological and socio-economic factors, as discussed later.

2.2 – Knowledge Discovery in Longitudinal Databases of Human Ageing

Knowledge Discovery in Databases (KDD) encompasses three broad tasks[18,51]: (a) data preparation, which includes e.g. data selection and data transformation; (b) the discovery of knowledge (or patterns) in databases, by using a data mining algorithm(s); (c) the evaluation and interpretation of the discovered knowledge in the context of the target problem domain.

In longitudinal databases of human ageing, each record represents an individual sampled from a population, and each individual is described by variables repeatedly measured across multiple time points (called 'waves'). Such longitudinal databases have some specific characteristics that should be considered in the KDD process, as discussed next.

When preparing longitudinal datasets of ageing for data mining, the temporal structure of the data should be considered. For example, missing value imputation for an age-dependent variable in a given wave can consider data from earlier and/or later waves. In the case of chronic age-related diseases, for instance, the presence of such a disease in a given wave can be used to infer the presence of that disease in a later wave, if the variable for the latter has missing values. In addition, longitudinal databases of human ageing typically have thousands of variables in each wave, and some individuals are not present in all waves. For instance, as individuals age they are increasingly

likely to leave the study, due to their death or frailty associated with their old age. In longitudinal data analyses, it is often desirable to create datasets where the same variables and individuals are kept throughout the waves.

In longitudinal ageing studies, the objective is to identify patterns expressing changes in the values of age-dependent variables or correlations between variables as a function of time. Such temporal patterns may be caused by several factors, such as seasonality (e.g. some age-related medical conditions may be more common in the winter) or the accumulated effect of the passage of time (e.g. the increase of hypertension risk as an individual ages). Conventional data mining (DM) algorithms ignore the temporal structure of longitudinal data, so longitudinal DM requires either that the data be carefully preprocessed to be suitable for conventional DM algorithms (a popular approach despite leading to loss of temporal information), or that new longitudinal DM algorithms be developed – possibly by adapting conventional DM algorithms.

The knowledge discovered in longitudinal databases of human ageing could potentially help to design new interventions against the harmful effects of the ageing process, improving human health, longevity, and productivity.

3 – A Taxonomy of Aspects Affecting Human Ageing

In order to evaluate the comprehensiveness of the main longitudinal databases of human ageing, we propose a taxonomy of the numerous aspects that influence the human ageing process. This taxonomy was based on an extensive review of the literature. The selection of the articles reviewed in this paper for creating the proposed taxonomy consisted of four phases, as follows.

First, we conducted searches for relevant articles in different academic reference repositories – namely, in PubMed and Springer repositories, and in the repositories of some of the most relevant journals that focus on human ageing from a multidisciplinary perspective (European Journal of Ageing, Journal of Aging Studies, Journal of Population Ageing, Ageing and Society). The searches considered articles published in the period 2004-2016, using as keywords "longitudinal database" *and* "ageing" (or "aging"). These searches returned about 500 articles. Second, because of the large volume of articles found, we performed part of the process of reduction proposed by Kitchenhan et al[36]. This reduction consisted in selecting the potentially relevant articles by reading the titles and abstracts of all articles found in the first search, and eliminating those unrelated to this research's objectives and scope, which reduced the number of articles to about 140. Third, we read in detail all those 140 articles, and then discarded many articles which were not relevant to our review, due to reasons such as not mentioning enough details about the ageing-related longitudinal variables analysed and not mentioning enough details about the data mining methods used. This further reduced the number of articles to 64 relevant ones. Fourth, these 64 articles were then very carefully reviewed and compared with each other, in order to identify ageing-related aspects across articles. In this fourth phase we have finally identified all aspects included in the proposed taxonomy, and have also identified the data source, geographical region and main data analysis technique applied in each reviewed research article.

Note that in this last phase it was also necessary to standardize the identified aspects' terminology, since there are discrepancies on how they were referred to in the reviewed articles, and some of the variables addressed in the reviewed articles had to be generalized, in order for them to suit the definition of some aspect in the taxonomy. Our review has identified in total 93 aspects, which have been organized into the taxonomy shown in Figure 2. Note also that some aspects may naturally have some overlap, as a result of the fact that they are described by relatively broad terms whose interpretation is to some extent subjective, given the nature of the concepts represented by some terms. In particular, broadly speaking, the taxonomy's terms referring to some social sciences or psychology concepts naturally tend to have a more subjective interpretation (and therefore are more prone to overlapping) than terms referring to biomedical or physical concepts. In

our review of the literature, we did our best to choose terms for the proposed taxonomy that best reflect the way the reviewed articles approached the studied concepts.

The taxonomy consists of eight broad dimensions (Physical Health, Quality of Life, Mental Health, Genetic Inheritance, Location, Personal, Social and Economic), each of which is further sub-divided into more specific aspects at different levels of granularity. Hence, the hierarchical nature of the taxonomy can help researchers to identify which types of ageing-related aspects (and at which level of abstraction) they should focus on, depending on their projects' goals. For example, if a research project is focused on discovering correlations between, say, data on the Psychological Healthy aspect (in the Mental Health dimension) and data on the Economic dimension, it could consider variables related to the psychological aspects Depression, Mental Disposition, Stress Level and Anxiety Level as well as variables related to Financial Situation, Occupation and Household (see Figure 2 for details).

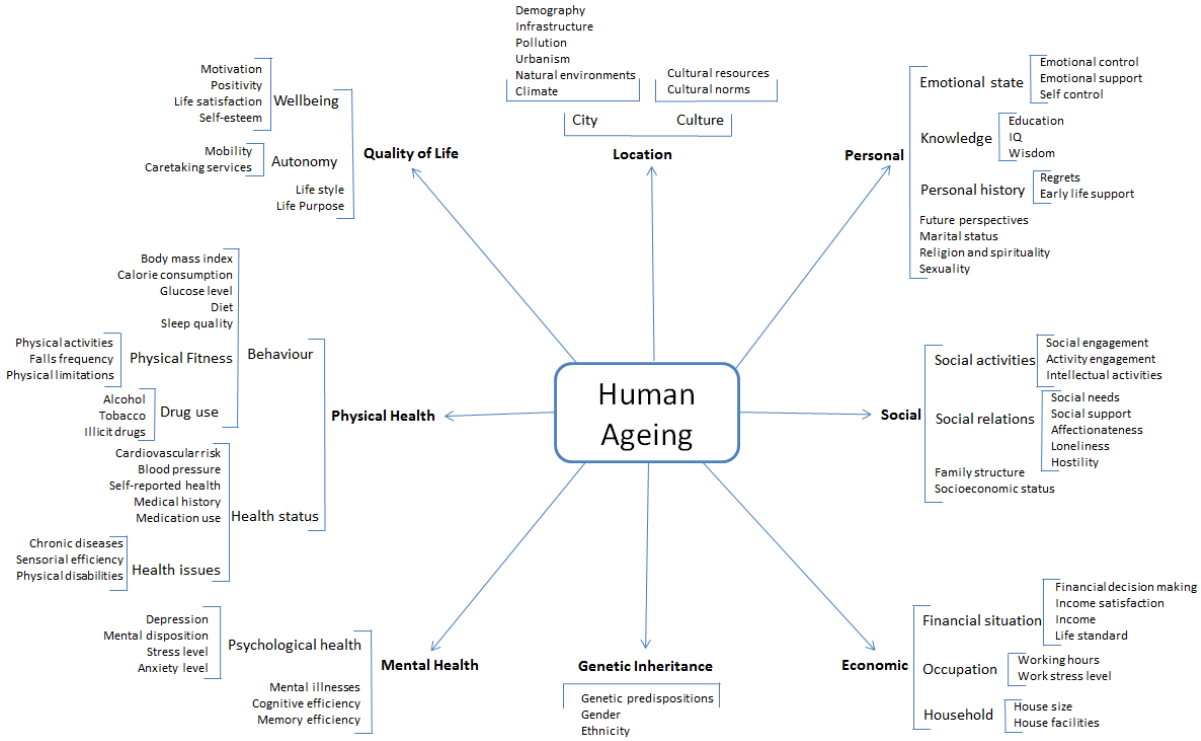

Figure 2: A taxonomy of aspects that influence human ageing.

Table 1 lists references for 63 of the 64 articles reviewed in this work, organized according to the dimensions addressed and analysis techniques used in each article (note that a reference can appear in more than one cells of the table). A single article[13] employed Generalized Linear Models, and we chose not to show it on this table to avoid the creation of another column in the table, due to space constraints, but it is included in the analysis nonetheless.

We briefly discuss next a couple of examples showing how data analysis methods can consider connections among different aspects of the proposed taxonomy or connections between aspects and other ageing-related variables. For a much more detailed review of such papers, a large table with detailed information on each reference, including the main ageing-related conclusions of the authors, is available in the first supplementary file (Table S1).

Table 1. Reference numbers of the reviewed articles, organized by dimensions of the proposed taxonomy and data analysis techniques.

DIMENSION	DATA ANALYSIS TECHNIQUE								
	Logistic Regression	Multilevel Logistic Regression	Multivariate Logistic Regression	Multilabel Logistic Regression	Linear Regression	Empirical Observations	Structural Equation Modelling	Cox Proportional Hazards	Analysis of Variance (ANOVA)
Physical Health	20, 21, 28, 37, 45, 50, 60, 75	38, 61	58, 71, 78		7, 8, 10, 26, 30, 39, 41, 62	69, 79	67, 73	40, 53, 54, 76	41, 65
Quality of Life	1, 20, 52, 60	24, 59, 63, 74		9, 70	10, 14, 26, 35, 62, 66, 77, 80	23, 34	77		65
Mental Health	52, 75	2, 38, 55, 61, 63	78		8, 10, 11, 19, 26, 30, 39, 41, 57, 77	27, 34	42, 56, 67, 73, 77	40, 33, 76	41, 65
Genetic Inheritance	3			70	11	79	67	33	
Location	3	63	71		62, 66	27	42		
Personal	21,28,37, 75	55, 74, 82	58	81	4, 10, 11, 14, 30, 41, 80, 57	27, 69	42, 67, 73	22	41
Social	21,52, 75	2, 55, 59, 74, 82	71	81	8, 11, 14, 26, 46, 66, 80	23, 34		22, 29, 33	
Economic	3	38, 61, 72	71	64	4, 7, 11, 19, 35, 39, 46	23, 27		53	

For example, the 'cognitive efficiency' aspect ('Mental Health' dimension) has been studied together with aspects like 'wellbeing'[2,77] ('Quality of Life' dimension), 'education'[57] ('Personal' dimension), and 'sensorial efficiency'[41] ('Physical Health' dimension). Moreover, 'physical and social activities' reportedly influenced positively the cognitive efficiency[8].

As another example, the aspect 'depression' ('Mental Health' dimension) has been associated with ageing-related variables like 'premature mortality'[76] and 'diabetes'[38,40] (the latter is included in the 'chronic disease' aspect in the 'Physical Health' dimension).

Figure 3 shows the proportion of reviewed articles using each of 9 broad types of data analysis techniques. Note that the reviewed articles have used in general classical statistical analysis techniques, and a preference for Linear and Logistic Regressions is evident in this Figure.

These classical statistical techniques have some limitations in the analysis of complex data, like longitudinal databases of human ageing. In particular, linear regression (in its various forms) and conventional correlation analysis-based techniques are usually parametric, making strong assumptions about the data distribution – such as assuming that the relationship between the dependent and independent variables is nearly linear, or assuming a normal distribution of the dependent variable.

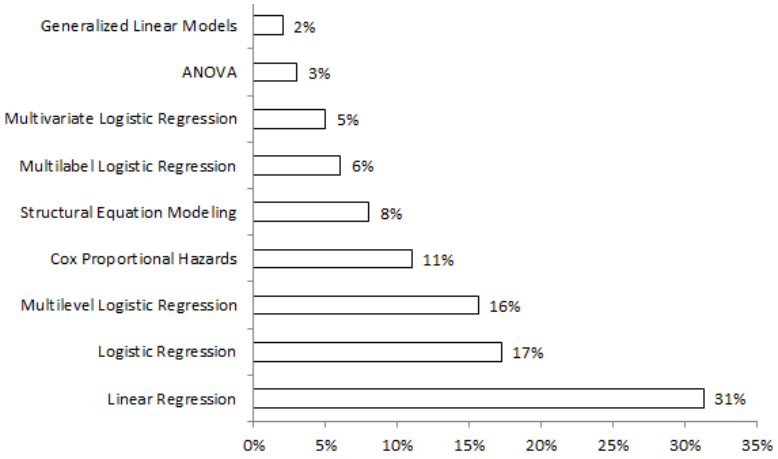

Figure 3: Percentage of reviewed articles that used each data analysis technique

When the dependent variable does not follow the normal distribution, Generalized Linear Models (GLMs)[47] can be used. As an extension of classical regression techniques, GLMs allow multivariate analysis with either continuous or nominal dependent variables, with an exponential or normal distribution. Note, however, that only one of the reviewed articles used GLM, namely, Choi et al.[13]..

Several studies addressed problems where the dependent variable is the time elapsed until an event (e.g. death). For instance, White[76] investigates how depression-related factors may shorten the lives of older adults. Several projects used the Cox Proportional Hazards Model [22,29,33,40,53,54], which is widely utilized in the analysis of survival rates. This model allows the identification of the independent variables that act more intensely when all variables are analysed as a set.

Another alternative to the limitations stated previously is Structural Equation Modelling (SEM)[25]. Briefly, SEM analyses the association between observable variables and latent variables – i.e. variables that are the product of relations between observable variables, but cannot be directly observed. Aspects such as wellbeing, that cannot be directly observed, may be considered latent variables. Based on this, Wilson[77] utilized SEM to assert that cognitive efficiency decline hinders wellbeing among older adults. SEM was also used in [42,56,67,73].

Among the articles included in our literature review, the proportions of articles addressing each of the 8 dimensions of the proposed taxonomy of ageing-related aspects are shown in Figure 4. Clearly, the 'Mental Health' and 'Physical Health' dimensions have been the most investigated in the reviewed articles.

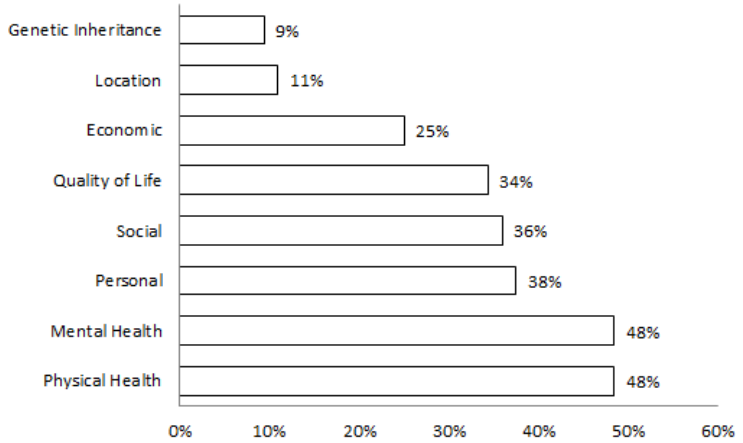

Figure 4: Percentage of articles analysing data from each dimension of the proposed taxonomy

4 – A Comprehensiveness Analysis of the Main Longitudinal Studies on Human Ageing

In this section we analyse the degree of comprehensiveness of the seven most prominent longitudinal studies on human ageing. By comprehensiveness we mean the extent to which they contain data about the different aspects identified in the proposed taxonomy (Figure 2). Table 2 presents an overview of the main characteristics of these longitudinal studies.

Table 2. Main characteristics of longitudinal databases of human ageing analysed in the reviewed articles.

DESCRIPTION	ELSA	SHARE	TILDA	CLHLS	WLS	KloSA	HRS
Number of aspects addressed (out of 93 in the taxonomy)	57	46	42	34	33	33	29
Average number of respondents per wave	12000	110000	8504	19000	10317	10000	26000
Number of waves published until June/2016	7	5	2	5	6	4	13
Geographical regions	United Kingdom	20 European countries + Israel	Ireland	China	Wiscosin, U.S.A.	South Korea	U.S.A.

ELSA – English Longitudinal Study of Ageing (site: http://www.elsa-project.ac.uk/)
SHARE - Survey of Health Ageing and Retirement in Europe (site: http://www.share-project.org/)
TILDA – Irish Longitudinal Study on Ageing (site: http://tilda.tcd.ie/)
CLHLS – Chinese Longitudinal Healthy Longevity Survey (site: http://centerforaging.duke.edu/chinese-longitudinal-healthy-longevity-survey)
WLS – The Wisconsin Longitudinal Study (site: http://www.ssc.wisc.edu/wlsresearch/)
HRS – Health and Retirement Study (site: http://hrsonline.isr.umich.edu/)
KloSA – Korean Longitudinal Study of Aging (site: http://survey.keis.or.kr/ENLCTGO01N.do?mnucd=cfsaklosa1)

The proportion of reviewed articles that used data from each longitudinal study is shown in Figure 5. Although the 'Others' category represents 44% of the reviewed articles, there are twenty databases in that category, and none of them has been used more frequently than the seven longitudinal databases shown in Table 2. In the first supplementary file (Table S1) we include the names of the databases used in each study, with a link to their website. The file also presents information on each article included in our revision, such as the main conclusions of the authors.

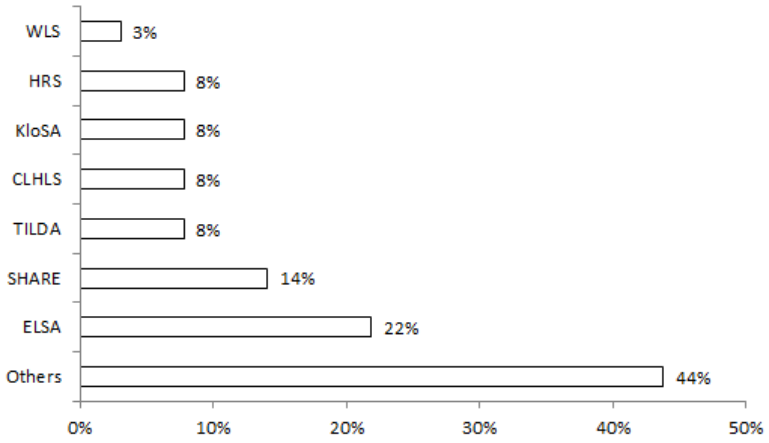

Figure 5: Percentage of articles using data from each database

The comprehensiveness analysis is presented in Table 3. Each row of the table refers to a different dimension (top-level aspect) in the taxonomy (Figure 2), and the columns refer to the

seven longitudinal studies on human ageing. The value in each cell indicates the comprehensiveness of the study for the corresponding dimension, represented by the percentage of aspects of that dimension addressed by the study. In our second supplementary file (Tables S2 through S9) we provide a more detailed analysis of comprehensiveness, with one table for each dimension, showing which aspects of that dimension are addressed by each study. To construct the table, the decision about exactly which aspects were addressed by each study required an interpretation of the documentation from each study's official website, since there are differences in nomenclature across studies, and several pieces of information are presented with little specificity. Hence, it is possible that an aspect not considered in our analysis is (directly or indirectly) addressed in the study's database.

Note that an aspect might be represented by multiple variables in a database. For instance, 'Physical Activities' is an aspect addressed by three variables in the ELSA database, representing the respondent's frequency of participation on light, moderate and intense physical activities.

Table 3. Analysis of the comprehensiveness of the main longitudinal databases of human ageing

DIMENSION	ELSA	SHARE	TILDA	CLHLS	WLS	KLoSA	HRS
Physical Health	**85.0%**	80.0%	70.0%	60.0%	15.0%	60.0%	50.0%
Quality of Life	**87.5%**	75.0%	**87.5%**	75.0%	37.5%	62.5%	37.5%
Mental Health	**85.7%**	57.1%	71.4%	57.1%	42.9%	42.9%	28.6%
Genetic Inheritance	**100%**	66.6%	66.6%	66.6%	**100.0%**	66.6%	**100%**
Location	62.5%	25.0%	25.0%	25.0%	**100.0%**	25.0%	25.0%
Personal	**50.0%**	33.3%	33.3%	16.7%	41.7%	25.0%	16.7%
Social	**80.0%**	70.0%	40.0%	40.0%	40.0%	30.0%	20.0%
Economic	**100.0%**	87.5%	75.0%	50.0%	87.5%	62.5%	**100.0%**

The studies in Table 3 have different levels of comprehensiveness across the dimensions of the proposed taxonomy, besides having different types of data and numbers of respondents, and covering diverse geographical regions (see Table 2). Hence, each of these studies has unique features that would make it more recommendable to certain types of research:

- **ELSA**: The English study established a quality standard referred to in most of the main studies on human ageing. Amongst the evaluated studies, ELSA stands out as the most comprehensive on most of the dimensions, except 'Location'. Furthermore, the study has a significant sample size, with about 12000 respondents per wave. Hence, ELSA can be recommended for most general purpose human ageing study projects.
- **SHARE**: This study is the most geographically distributed, being conducted on 20 European countries and Israel. It also has by far the largest sample of respondents (see Table 2). Furthermore, it is the second most comprehensive study overall, in Table 3.
- **TILDA**: The Irish longitudinal study has a good comprehensiveness on every dimension, and a stable database, due to using a model similar to ELSA's. Because it was most recently started (the first wave was conducted in 2009), TILDA has adhered to a more mature ELSA model, without the need to pass through the several adaptations that ELSA had to. Thus, the data preparation phase of a project using the TILDA database should, in theory, need fewer adaptations and concerns with the data format.

- **CLHLS**: This study focuses on behavioural analysis, aiming to identify ageing-related risk behaviours. The Chinese database has the greatest number of centenarians (14290) and a considerable number of nonagenarians (18910) and octogenarians (14416). Hence, studies that aim to find answers to the very high longevity of individuals might benefit from its data.
- **WLS**: The Winscosin study has the oldest data available, some dating back to the 50's. Hence, an analysis of its data can provide insights on the effect of the passage of greater periods of time than other studies. Also, it is focused on a single city, which allows it to fully cover the 'Location' dimension and have more specific variables for that dimension.
- **KLoSA:** Another database that follows the ELSA model, KloSA is limited to South Korea respondents, which, like the Chinese respondents in CLHLS, have specific social and cultural characteristics. Because they have oriental respondents, KloSA and CLHLS are recommended for projects that compare environmental differences between cultures, possibly comparing results from these studies with the ones from studies that focus on occidental cultures.
- **HRS**: Many respondents of this study (about 20000 individuals) had their genetic codes mapped and added to the study's database. Hence, projects may use this database for analyses that aim to correlate genetic influences with the environmental characteristics of the participants. In addition, the study currently has 13 published waves, more than every other study in Table 2, which allows a more extensive longitudinal analysis.

5 – Data Analysis Techniques Applied to Longitudinal Databases of Ageing

As previously mentioned, the data analysis techniques most commonly used to analyse longitudinal databases of human ageing are classical statistical techniques such as linear and logistic regression. The articles selected in our review utilized data from several longitudinal studies, and applied different data analysis techniques, as presented in Table 4.

Table 4. Reference numbers of the reviewed articles, organized by data analysis technique and longitudinal database of human ageing

LONGITUDINAL DATABASE		ELSA	SHARE	TILDA	CLHLS	WLS	KloSA	HRS
DATA ANALYSIS TECHNIQUE	Logistic Regression	20, 21, 28, 37, 52, 60		3, 45	75			50
	Multilevel Logistic Regression	2, 59, 61, 72	72, 74	55	82		38	72
	Multivariate Logistic Regression	71			78			
	Multilabel Logistic Regression		70		81			
	Linear Regression	62, 80	10, 19, 57	46			35	11, 35
	Generalized Linear Model						13	
	Cox Proportional Hazards	76	22, 53			29	33	
	Analysis of Variance (ANOVA)		65					

Our literature review brought up an interesting issue: the lack of published articles analysing longitudinal databases of ageing with modern data mining algorithms like support vector machines (SVMs) and random forests (RFs). There are SVMs and RFs designed to cope with longitudinal data, but such algorithms have not been applied yet to the major longitudinal databases of ageing

discussed in this review, possibly due to the lack of awareness of the data mining community about the availability and importance of such databases.

As related work, we briefly mention three examples of SVMs applied to longitudinal data about a specific disease or health problem (rather than about the ageing process as a whole). First, Minhas[48] employs Support Vector Machine (SVM) classifiers to predict early Alzheimer's disease. Second, Bellazi[5] employs SVM classifiers to analyse longitudinal data, aiming to assess the clinical performance of haemodialysis services. As the third example, Du[16] uses SVM for predicting Amyotrophic Lateral Sclerosis (ALS) score. All three articles concluded that the use of SVM was successful.

6 - Conclusion

Longitudinal studies of human ageing gather information about thousands of individuals, collected over many years. The databases created by such studies contain a large number of variables, of diverse types, providing data about a number of different aspects affecting human ageing. In addition, such studies are conducted around the world, representing populations with diverse social, economic and cultural characteristics. Furthermore, the data stored in such longitudinal databases of ageing is often freely available for researchers.

Hence, the data contained in these databases is a very useful source of information for research on human ageing. In particular, data analysis techniques from statistics and data mining can be used to extract knowledge or patterns about the complex process of human ageing.

In order to support research based on longitudinal databases, we performed an extensive review of the literature on longitudinal studies of human ageing. Using the methodology described in Section 3, a large number of articles on the theme (about 500) were obtained. However, only 64 of them were considered relevant based on the inclusion and exclusion criteria adopted. As a limitation of this work, it is possible that the process of eliminating some papers reading only their title and abstract, although it is widely used, has resulted in discarding a few papers that would be found relevant if we had read all full papers. However, it would not be feasible or cost-effective to read 500 articles in detail, and in general the titles and abstracts of the discarded papers clearly indicated that they were very unlikely to be relevant for our review. This trade-off of analysing a large amount of papers in a reasonable amount of time is well known in information retrieval researches.

As for contributions, first, we proposed a taxonomy of the main aspects affecting human ageing, as shown in Figure 2. Second, as a summary of our literature review, we presented tables which categorize the reviewed articles according to the data analysis techniques they used and the broad ageing-related aspect that they investigate (Table 1) or the major longitudinal datasets that they analysed (Table 4). Third, we identified the seven most prominent longitudinal studies of human ageing around the world, and for each study we summarized its main characteristics (Table 2) and analysed its degree of comprehensiveness, that is, the extent to which they contain data about the different ageing-related aspects identified in the proposed taxonomy (Table 3).

We also noted that the data analysis techniques used so far to analyse longitudinal databases of human ageing have been mainly classical statistical techniques, often limited by parametric and linearity assumptions. Actually, to the best of our knowledge, there is no work applying more modern data mining methods to such longitudinal databases of human ageing, although these methods have been applied to longitudinal data about a specific disease or health condition (some examples were briefly mentioned in Section 5). The study of human ageing could greatly benefit from different forms of non-parametric and non-linear data analyses, such as modern data mining algorithms developed to deal with longitudinal data.

We hope that this review can help other researchers to apply statistical and data mining methods to longitudinal data on human ageing, contributing to the discovery of more knowledge in

this area. Note that this is a very important research area, since the proportion of the world population becoming old or very old is expected to substantially increase in the coming decades, with large socio-economic and health implications, as discussed earlier.

Acknowledgement

This work was conducted during a scholarship supported by the International Cooperation Program CAPES/COFECUB at the PUC-Minas University. Financed by CAPES – Brazilian Federal Agency for Support and Evaluation of Graduate Education within the Ministry of Education of Brazil.

References

1. Aartsen, M., Jylhä, M. Onset of loneliness in older adults: results of a 28 year prospective study. *European Journal of Ageing* 2011, 8:31–38.
2. Allerhand, M., Gale, C. R., and Deary, I. J. The dynamic relationship between cognitive function and positive well-being in older people: A prospective study using the English Longitudinal Study of Aging. *Psychology and Aging* 2014, 29:306-318.
3. Barrett, A., Mosca, I. Early-life Causes and Later-life Consequences of Migration: Evidence from Older Irish Adults. *Journal of Population Ageing* 2013, 6:29–45.
4. Behrman, J., Kohler, H.-P., Jensen, V., Pedersen, D., Petersen, I., Bingley, P., Christensen, K. Does More Schooling Reduce Hospitalization and Delay Mortality? New Evidence Based on Danish Twins. *Demography* 2011, 48:1347–1375.
5. Bellazi, R., Larizza, C., Magni, P., Bellazi, B., Temporal data mining for the quality assessment of hemodialysis services. *Artificial Intelligence in Medicine*, 2015, 34.1:25-39
6. Bergeman C. Aging: Genetic and environmental influences. *SAGE Publications* 1997.
7. Berthoud, R., Blekesaune, M. and Hancock, R., Ageing, income and life standards: evidence from the British Household Panel Survey. *Ageing and Society* 2009, 29:1105-1122
8. Bielak, A. M., Anstey, K. J., Gerstof, D., and Luscz, M. A. Longitudinal associations between activity and cognition vary by age, activity type, and cognitive domain. *Psychology and Aging* 2014, 29:863-872.
9. Blekesaune, M., Skirbekk, V. Can personality predict retirement behaviour? A longitudinal analysis combining survey and register data from Norway. *European Journal of Ageing* 2012, 9:199–206.
10. Bourassa, K. J., Memel, M.,Woolverton, C., and Sbarra, D. A. A dyadic approach to health, cognition, and quality of life in aging adults. *Psychology and Aging* 2015, 30:449–461.
11. Cacioppo, J. T., Waite, L. J., Hawkley, L. C. and Thisted, R. A. Loneliness as a specific risk factor for depressive symptoms: Cross-sectional and longitudinal analyses. *Psychology and Aging* 2006, 21:140-151
12. Chen, T., Zeng, D., Wang, Y. Multiple kernel learning with random effects for predicting longitudinal outcomes and data integration. *Biometrics* 2015, 71:918-928.
13. Choi, Y., Park, E.-C., Kim, J.-H., Yoo, K.-B., Choi, J.-W., and Lee, K.-S. A change in social activity and depression amongst Koreans aged 45 years and more: analysis of the Korean Longitudinal Study of Aging (2006-2010). *International Psychogeriatrics* 2015, 27:629–637.
14. Cowlishaw, S., Niele, S., Teshuva, K., Browning, C., and H., K. Older adults spirituality and life satisfaction: a longitudinal test of social support and sense of coherence as mediating mechanisms. *Ageing and Society* 2013, 33:1243–1262.
15. Diggle, P. J., Heagertly, P., Liang, K. and Zeger, S., Analysis of Longitudinal Data - Second Edition. *Oxford University Press* 2013.
16. Du, W., Cheung, H., Johnson, C.A. A longitudinal support vector regression for prediction of ALS score. *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 2015, 1586-1590.
17. Epel, E. and Lithgow G.J. Stress Biology and aging mechanism: toward understanding the deep connection between adaptation to stress and longevity. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 2014; 69(S1):S10-S16.
18. Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. From Data Mining to Knowledge Discovery in Databases. *AI Magazine, v. 17, n. 3, p. 37* 1996.
19. Fonseca, R., Kapteyn, A., Lee, J., Zamarro, G., Feeney, K. A Longitudinal Study of Well-Being of Older Europeans: Does Retirement Matter? *Journal of Population Ageing* 2014, 7:21–41.
20. Frisher, M., Mendonça, M., Shelton, N., Pikhart, H, de Oliveira, C. and Holdsworth, C. Is alcohol consumption in older adults associated with poor self-rated health? Cross-sectional and longitudinal analyses from the English Longitudinal Study of Ageing. *BMC public health, v. 15, n. 1, p. 1* 2015.

21. Gale, C. R., Cooper, C., Deary, I. J. and Sayer, A. A. Psychological well-being and incident frailty in men and women: the English Longitudinal Study of Ageing. *Psychological medicine* 2014, 44:697-706.

22. Gumà, J., Cámara, A., Treviño, R. The relationship between health and partnership history in adulthood: insights through retrospective information from Europeans aged 50 and over. *European Journal of Ageing* 2014; 12:71–79.

23. Hauser, R., Weir, D. Recent developments in longitudinal studies of aging in the United States. *Demography* 2010; 47:S111–S130.

24. Herrera Ponce, M., Barros Lezaeta, C., Fernández Lorca, M. Predictors of Quality of Life in Old Age: A Multivariate Study in Chile. *Journal of Population Ageing* 2011; 4:121–139.

25. Hoyle, R. H. (Ed.). Structural equation modeling: Concepts, issues, and applications. *Sage Publications* 1995.

26. Hsu, Y.-C., Chiu, C.-J., Wray, L.A., Beverly, E.A., and Tseng, S.-P. Impact of traditional Chinese medicine on age trajectories of health: evidence from the Taiwan Longitudinal Study on Aging. *Journal of the American Geriatrics Society* 2015, 63:351–357.

27. Hudson, E., Barrett, A.Peer Groups, Employment Status and Depressive Symptoms Amongst Older Adults in Ireland. *Journal of Population Ageing* 2014, 7:43–54.

28. Jackson, S.E., Steptoe, A., and Wardle, J. The influence of partner's behavior on health behavior change: the English Longitudinal Study of Ageing. *JAMA Internal Medicine* 2015, 175:385–392.

29. Jin, L., Elwert, F., Freese, J., Christakis, N. Preliminary evidence regarding the hypothesis that the sex ratio at sexual maturity may affect longevity in men. *Demography* 2010, 47:579–586.

30. Johnson, W., McGue, M., Deary, I. J., and Christensen, K. Genetic and environmental transactions linking cognitive ability, physical fitness, and education in late life. *Psychology and Aging* 2009; 24:48–62.

31. Kaiser, Angelika. A Review of Longitudinal Datasets on Ageing. *Journal of Population Ageing*, 2013, 6.1-2:5-27.

32. Kantardzic, M., Data Mining: Concepts, Models, Methods and Algorithms, *John Wiley & Sons* 2002.

33. Kim, J.-H., Lee, S.G., Shin, J., Choi, Y., and Park, E.-C. The effect of offspring on depressive disorder amongst old adults: Evidence from the Korean Longitudinal Study of Aging from 2006 to 2012. *Archives of Gerontology and Geriatrics* 2015, 61:351–362.

34. Kim, J-H, Lee, S. G., Shin, J, Cho, K. H., Choi, J. W. and Park, E. C. Effects of number and gender of offspring on quality of life among older adults: evidence from the Korean Longitudinal Study of Aging, 2006–2012. *BMJ open*, v. 5, n. 6, p. e007346, 2015.

35. Kim, S., Sargent-Cox, K. A., French, D. J., Kendig, H., and Anstey, K. J. Cross-national insights into the relationship between wealth and wellbeing: a comparison between Australia, the United States of America and South Korea. *Ageing and Society* 2012; 32:41–59.

36. Kitchenhan, Barbara et al. "Systematic Literature Reviews in Software Engeneering – a systematic literature review". *Information and Software Technology* 2009, 51.1:7-15.

37. Lee, D. M., Nazroo, J., O'Connor, D. B., Blake, M. and Pendleton, N. Sexual health and well-being among older men and women in England: findings from the English Longitudinal Study of Ageing. *Archives of sexual behaviour* 2016, 45:133-144.

38. Lee, Hyun Kyung; Lee, Seung Hee. Depression, diabetes, and healthcare utilization: results from the Korean longitudinal study of aging (KLoSA). *Iranian journal of public health, v. 43, n. 1, p. 6* 2014.

39. Lee, J., Shih, R., Feeney, K., and Langa, K.M. Gender disparity in late-life cognitive functioning in India: findings from the longitudinal aging study in India. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 2014, 69:603–611.

40. Limongi, F., Noale, M., Crepaldi, G., and Maggi, S. (2014). Prevalence of diabetes and depressive symptomatology and their effect on mortality risk in elderly Italians: The Italian Longitudinal Study on Aging. *Diabetes & Metabolism* 2014, 40:373–378.

41. Lindenberger, U. and Ghisletta, P. Cognitive and sensory declines in old age: Gauging the evidence for a common cause. *Psychology and Aging* 2009, 24:1–16.

42. Luszcz, M.A., Anstey, K.J., and Ghisletta, P. (2015). Subjective Beliefs, Memory and Functional Health: Change and Associations over 12 Years in the Australian Longitudinal Study of Ageing. *Gerontology* 2015, 61:241–250.

43. Lutz, W., Sanderson, W., Scherbov, S. The coming acceleration of global population ageing. *Nature* 2008, 451:716-719.

44. Malloy-Diniz, L. F., Fuentes, D., & Cosenza, R. M. Neuropsicologia do envelhecimento: uma abordagem multidimensional. *Artmed Editora* 2013.

45. Mc Hugh, S., O'Neill, C., Browne, J., and Kearney, P.M. Body mass index and health service utilisation in the older population: results from The Irish Longitudinal Study on Ageing. *Age & Ageing* 2015, 44:428–434.
46. McCrory, C., Finucane, C., O'Hare, C., Frewen, J, Nolan, H., Layte, R., Kearney, P. M. and Kenny, R. A. Social disadvantage and social isolation are associated with a higher resting heart rate: Evidence from The Irish Longitudinal Study on ageing. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 2015, 71:463-473.
47. McCullagh, P., & Nelder, J. A. Generalized linear models (Vol. 37). *CRC press* 1989.

48. Minhas, Sidra et al. Early Alzheimer's Disease Prediction in Machine Learning Setup: Empirical Analysis with Missing Value Computation. *Conference on Intelligent Data Engeneering and Automated Learning.* Springer International Publishing, 2015, 424-432.
49. Mock, S. E. and Eibach, R. P. Aging attitudes moderate the effect of subjective age on psychological well-being: Evidence from a 10-year longitudinal study. *Psychology and Aging* 2011, 26:979–986.
50. Monteverde, M., Noronha, K., Palloni, A., Novak, B. Obesity and excess mortality amongst the elderly in the United States and Mexico. *Demography* 2010, 47:79–96.
51. Pyle, D. Data preparation for data mining. Vol. 1. *Morgan Kaufmann* 1999.
52. Rafnsson, Snorri Bjorn; Shankar, Aparna; Steptoe, Andrew. Informal caregiving transitions, subjective well-being and depressed mood: Findings from the English Longitudinal Study of Ageing. *Aging & Mental Health* 2015, p. 1-9.
53. Romero-Ortuno, Roman; Soraghan, Christopher. A Frailty Instrument for primary care for those aged 75 years or more: findings from the Survey of Health, Ageing and Retirement in Europe, a longitudinal population-based cohort study (SHARE-FI75+). *BMJ open, v. 4, n. 12, p. e006645* 2014.
54. Sanchez-Santos, M., Zunzunegui, M., Otero-Puime, A., Cañas, R., Casado-Collado, A. Self-rated health and mortality risk in relation to gender and education: a time-dependent covariate analysis. *European Journal of Ageing* 2011; 8:281–289.
55. Santini, Z.I., Koyanagi, A., Tyrovolas, S., and Haro, J.M. The association of relationship quality and social networks with depression, anxiety, and suicidal ideation amongst older married adults: Findings from a cross-sectional analysis of the Irish Longitudinal Study on Ageing (TILDA). *Journal of Affective Disorders* 2015, 179:134–141.
56. Schmiedek, F. and Li, S. Toward an alternative representation for disentangling age-associated differences in general and specific cognitive abilities. *Psychology and Aging* 2004, 19:40–56.
57. Schneeweis, N., Skirbekk, V., Winter-Ebmer, R. Does Education Improve Cognitive Performance Four Decades After School Completion? *Demography* 2014, 51:619–643.
58. Schoenmakers, E. C., Van Tilburg, T. G., and Fokkema, T. Awareness of risk factors for loneliness amongst third agers. *Ageing and Society* 2014, 34:1035–1051.
59. Shankar, A., Rafnsson, S.B., and Steptoe, A. Longitudinal associations between social connections and subjective wellbeing in the English Longitudinal Study of Ageing. *Psychology & Health* 2015, 30:686–698.
60. Smith, L., and Hamer, M. (2014). Television viewing time and risk of incident diabetes mellitus: the English Longitudinal Study of Ageing. *Diabetic Medicine* 2014, 31:1572–1576.
61. Smith, L., Gardner, B., Fisher, A. and Hammer, M. Patterns and correlates of physical activity behaviour over 10 years in older adults: prospective analyses from the English Longitudinal Study of Ageing. *BMJ open, v. 5, n. 4, p. e007423* 2015.
62. Smith, L.; Fisher, A.; Hamer, M. Television viewing time and risk of incident obesity and central obesity: the English longitudinal study of ageing. *BMC Obesity, v. 2, n. 1, p. 1* 2015.
63. Solfrizzi, V., Panza, F., Imbimbo, B.P., D'Introno, A., Galluzzo, L., Gandin, C., Misciagna, G., Guerra, V., Osella, A., Baldereschi, M., et al. Coffee Consumption Habits and the Risk of Mild Cognitive Impairment: The Italian Longitudinal Study on Aging. *Journal of Alzheimer's Disease* 2015, 47:889–899.
64. Solinge, H. Who opts for self-employment after retirement? A longitudinal study in the Netherlands. *European Journal of Ageing* 2013, 11:261-272.
65. Thøgersen-Ntoumani, C., Barkoukis, V., Grano, C., Lucidi, F., Lindwall, M., Liukkonen, J., Raudsepp, L., Young, W. Health and well-being profiles of older European adults. *European Journal of Ageing* 2011, 8:75–85.
66. Tomaszewski, W. Living Environment, Social Participation and Wellbeing in Older Age: The Relevance of Housing and Local Area Disadvantage. *Journal of Population Ageing* 2013, 6:119–156.
67. Tucker-Drob, E. M., Briley, D. A., Starr, J. M., and Deary, I. J. Structure and correlates of cognitive aging in a narrow age cohort. *Psychology and Aging* 2014, 29:236-249.

68. UNDESA (United Nations Department of Economic and Social Affairs) Population Division. "World Population Prospects: The 2015 Revision, Key Findings and Advance Tables." *Working Paper ESA/P/WP.241* 2015.
69. Van Nes, F., Jonsson, H., Abma, T., Deeg, D. Changing everyday activities of couples in late life: Converging and keeping up. *Journal of Aging Studies* 2013, 27:82–91.
70. Verropoulou, G. Determinants of change in self-rated health amongst older adults in Europe: a longitudinal perspective based on SHARE data. *European Journal of Ageing* 2012, 9:305–318.
71. Vlachantoni, A., Shaw, R. J., Evandrou, M., and Falkingham, J. The determinants of receiving social care in later life in england. *Ageing and Society* 2015, 35:321–345.
72. Vries, R., Blane, D., Netuveli, G. Long-term exposure to income inequality: implications for physical functioning at older ages. *European Journal of Ageing* 2014, 11:19–29.
73. Wahlin, A., deFrias, C. M., MacDonald, S. W. S., and Nilsson, L. How do health and biological age influence chronological age and sex differences in cognitive aging: Moderating, mediating, or both? *Psychology and Aging* 2006, 21:318–332.
74. Wahrendorf, M., Siegrist, J. Are changes in productive activities of older people associated with changes in their well-being? Results of a longitudinal European study. *European Journal of Ageing* 2010, 7:59–68.
75. Wen, M., Gu, D. The Effects of Childhood, Adult, and Community Socioeconomic Conditions on Health and Mortality amongst Older Adults in China. *Demography* 2011, 48:153–181.
76. White, J., Zaninotto, P., Walters, K., Kivimaki, M., Demakakos, P., Shankar, A., Kumari, M., Gallacher, J., and Batty, G.D. Severity of depressive symptoms as a predictor of mortality: the English longitudinal study of ageing. *Psychological Medicine* 2015, 45:2771–2779.
77. Wilson, R. S., Boyle, P. A., Segawa, E., Yu, L., Benegy, C. T., Anagnos, S. E., and Bennett, D. A. The influence of cognitive decline on well-being in old age. *Psychology and Aging* 2013, 28:304–313.
78. Wu, Z. and Schimmele, C. M. Psychological disposition and self-reported health amongst the 'oldest-old' in China. *Ageing and Society* 2006, 26:135–151.
79. Yong, V., Gu, D., Chen, M., Saito, Y. Expected Lifetime with and without Cataract amongst Older Adults in China. *Journal of Population Ageing* 2011, 4:65–79.
80. Zaninotto, P., Breeze, E., McMunn, A., Nazroo, J. Socially Productive Activities, Reciprocity and Well-Being in Early Old Age: Gender-Specific Results from the English Longitudinal Study of Ageing (ELSA). *Journal of Population Ageing* 2013, 6: 47–57.
81. Zhu, H., Gu, D. The Protective Effect of Marriage on Health and Survival: Does It Persist at Oldest-Old Ages? *Journal of Population Ageing* 2010, 3:161–182.
82. Zimmer, Z., Korinek, K. Shifting coresidence near the end of life: Comparing decedents and survivors of a follow-up study in China. *Demography* 2010, 47:537–554.