Supplementary File 2

2. Data and Methodology

2.1 Data

Longitudinal Ageing Study in India (LASI), Main Wave I, 2017-18, is a longitudinal survey that provides data on many diseases for 73,396 individuals aged 45 years and above and their spouses irrespective of age by their socioeconomic and demographic details in the reference year 2017-18. It covers 28 states and 8 UTs of India. The LASI is considered a cross-sectional survey for analytical purposes because it is yet a first wave. LASI has adopted a multi-stage stratified area probability cluster sampling design, with three stages in the rural and four stages in the urban areas, respectively. Firstly, a primary sampling unit (PSU) was selected from each state/union territory (UT), followed by a village (from rural) or ward (from urban) area in the second stage stratum(1, 2). Secondly, rural households were selected from these selected villages in the second stage stratum. Besides, census enumeration blocks (CEBs) were selected randomly from each urban area. Lastly, urban households were chosen from the selected CEBs (International Institute for Population Sciences, 2020). The study used information on individual and biomarker datasets.

2.2 Methodology

We retrieved data on thirty-six diseases and conditions, along with their prevalence, from the LASI survey. The ICD (2010)(3) classification of these thirty-six diseases and conditions are provided in Table 1. These thirty-six diseases and conditions were further classified in binary forms (absent and present). A morbidity score was generated and categorised into (1) no morbidity (individuals with no diseases or conditions), (2) single morbidity (individuals with exactly one diseases or conditions) and (3) multimorbidity (combinations of two or more diseases or conditions). We focused on heart diseases (life-threatening diseases) to showcase the repercussions of co-multimorbidity, diseases and conditions co-occurrences on sentinel conditions. We have three major analyses incorporated in this study.

(1) We first examined the age-specific patterns of the diseases and conditions. Subsequently, a Chi-square test was employed to assess the distribution of morbidity across contextual factors (see Sections 3.3.1 and 2.2.2).
(2) We constructed the undirected network (see section 3.3.2) to understand the diseases and conditions interdependence and interactions for connectedness.
(3) We employed Classification and Regression Tree (CART) modelling (see Section 3.3.3) to examine the conditional role of diseases or conditions co-occurrences, along with contextual factors, in shaping the risk of heart diseases.

Also, the LASI survey provides the onset of only eleven diseases and conditions which are hypertension or high blood pressure, diabetes or high blood sugar, cancer or a malignant tumour, chronic lung disease, chronic heart disease, heart failure or other heart problems, stroke, arthritis or rheumatism, osteoporosis, neurological or psychiatric problems, and high cholesterol, which

was analysed for a corollary analysis (see supplementary file 1-Table 4). 2.2.1 Outcome or Dependent Variable

The outcome variable is the conditional probability of a heart diseases (consisting of any of heart attack, heart blockage, heart failure, arrhythmias heart, rheumatic heart, heart congenital or structural disorder, and other heart diseases) in the presence of co-occurrences of diseases/conditions and socioeconomic, demographic, and risk factors. Heart diseases are reported on self-reported data and documented diagnoses from the LASI dataset, obtained through participant responses to survey questions such as “Has any health professional ever diagnosed you with the following chronic conditions or diseases?” and “In the past two years, have you had any of the following diseases?”

2.2.2 Covariates or Contextual Factors or Independent Variables

We have considered socioeconomic, demographic, behavioural, and lifestyle risk factors to explore the differentials in multimorbidity. We have adopted Anderson's healthcare utilization framework (4, 5) for the selection and categorising of the independent variables in three groups:

i. *Predisposing factors* were Age, Sex (Male/Female), Caste groups (SCs/STs, OBCs, Others), Religion (Hindu, Muslim, Christian & Others), Marital status (Currently Married, Widowed, Divorced/Separated), Living arrangement (Alone, Spouse, Other), Impairment (No/ Yes).
ii. *Enabling factors* were Level of education (No schooling, < 5 Years,5-9 years, 10 + years), Wealth index/MPCE quintiles (poorest, poorer, middle, richer, richest), Residence (Rural, Urban), Regions (North, Central, East, Northeast, West, South), Working status (never worked, currently working, currently not working), Childhood health (Very good, Good, Fair, Poor, Very Poor), Self-rated health (Good, Moderate, Poor), Physical activity (Everyday, Weekly, Casual).
iii. *Behavioural and lifestyle* factors were Tobacco consumption (Lifetime abstainer, Smokes tobacco, Smokeless tobacco, Both), Alcohol consumption (Lifetime abstainer, Infrequent non-heavy drinker, Frequent non-heavy drinker, Heavy episodic drinker), BMI (Underweight, Normal, Overweight, Obese).
iv. Health status factors were comprised functional health indicators (ADL and IADL limitations)

3.3 Statistical Analyses

3.3.1 Frequency distribution and Chi-square test

We calculated the prevalence (per hundred) of thirty-six diagnosed diseases and conditions reported by respondents in the LASI survey, along with the proportions of individuals with zero morbidity (no diseases or conditions), single morbidity (one diseases or conditions), and multimorbidity (two or more diseases or conditions). We tabulated the morbidity status by socioeconomic, demographic, and lifestyle factors. We hypothesized that the distribution of no morbidity, single, and multimorbidity may be similar or different by the covariates; for this purpose, the Chi-square test was applied to test the null hypothesis of the same distribution of morbidity statuses.

3.3.2 Undirected Network Analysis

We constructed an undirected network(6) of thirty-six diseases and conditions to present their complex interactions, i.e. interdependence and simultaneity, in the Indian olde adults population. The network analysis of diseases and conditions comprises nodes and edges and connections between them. Each node represents a disease or condition, and its frequencies of co-occurrences with other nodes or the connection between two nodes are edges. The edge manifests the co-occurrence of diseases or conditions in the network, and the weight of edges represents the strength of association between the two nodes. Many edges connected with a node represent the connections between diseases or conditions in the network. A node with many edges characterised by the long or short length may be crucial to the disease network. The impact of a node and its edges in a disease network is measured in terms of degree of centrality, betweenness centrality and closeness centrality.

Also, the number of nodes (diseases or conditions), the number of edges (co-occurrences of diseases and conditions), network density (the extent to which diseases are connected) and network diameter (maximum distance between any two nodes) were computed to evaluate the connectedness, cohesiveness, and sparsity of the network. Connections are defined according to the criteria of co-occurrences to a greater degree than expected by the prevalence of diseases. For these criteria, thresholds were used in the association measures (p-value < 0.05). We applied these filters because the large sample size analysed would generate significant associations even for very small magnitudes.

3.3.2.1 Nodes attributes and network parameters

We calculated the number of nodes, edges, diameter, density, and centralities. The number of nodes is considered the number of units in the network, and the number of edges is the number of links connecting the nodes. Network diameter is the maximum distance between any two nodes in a network, which is calculated as

$$Network\ diameter = \{d_{ij} | i, j\ \in N\} \qquad \text{..... Eq. (1)}$$

where N is the set of nodes in the network.

The network density is the number of connections between nodes divided by the combinations of possible ties or connections in a network. The value of network density ranges between 0 and 1. It is calculated as

$$Density = \frac{\sum_i \quad \sum_j \quad m_{ij}}{n(n-1)} \qquad \text{..... Eq. (2)}$$

where $M = [m_{ij}]$ is the adjacency n×n matrix.

3.3.2.2 Measures of centrality

The degree of centrality, betweenness centrality, and closeness centrality measures were used to analyse the impact and influence of a disease or conditions with other diseases or conditions in the network. The degree of centrality measure describes the number of connections linked to nodes. In other terms, degree centrality measures the number of nodes in neighbourhoods.

***Degree*:** It measures the connectivity of the node as the number of connectedness to node 'i'. It is computed as

$$K_i = k_{in}(i) + k_{out}(i) \quad \text{..... Eq. (3)}$$

where, K_i is the total number of connections of a node 'i' with other nodes, $k_{in}(i)$ is the number of ties directed inwards, $k_{out}(i)$ number of ties directed outwards in a network.

***Degree of centrality*:** It calculates the fraction of edges that a particular node 'i' has compared to the total number of nodes in the network. Nodes with a higher degree of centrality are more connected with other nodes in a network. The measure is given by

$$C_d(i) = \frac{number\ of\ edges\ connected\ to\ node\ i}{total\ number\ of\ nodes - 1} \quad \text{..... Eq. (4).}$$

***Betweenness centrality*:** The betweenness centrality measures how often a node lies on the shortest path between two nodes. It signifies the node that acts as a bridge along the shortest paths (edges) connecting two other nodes within a network. It sums the fraction of shortest paths between all pairs of nodes '*s*' and 't' that pass-through node 'i'. It is expressed as

$$C_b(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}} \quad \text{..... Eq. (5)}$$

where, σ_{st} is the total number of shortest paths from the node *'s'* to node *'t'*; $\sigma_{st}(i)$ is the number of those paths that pass-through node '*i*'.

***Closeness centrality*:** It assesses how close a node is to all other nodes in the network. It is the reciprocal of the sum of the shortest paths from the node 'i' to all other nodes. Nodes with high closeness centrality are close to many other nodes in the network.

$$C_c(i) = \frac{1}{\sum_u shortest\ path(i,j)} \quad \text{..... Eq. (6).}$$

3.3.3 Classification and Regression Tree (CART) Analysis

The Classification and Regression Tree (CART) model(7) was performed to showcase the co-occurrences of diseases or conditions explaining the risk of heart diseases. It provides the conditional probability of heart disease; the probability is conditioned upon the presence of co-occurrences of diseases as well as contextual factors.

The CART model is a machine learning, non-parametric statistical method requiring no distributional assumptions on explanatory variables. This model is a robust, nonlinear method that optimizes predictive accuracy by fitting an ensemble of decision trees to stabilize model estimates. This method uses a set of bootstrap samples, growing an independent decision tree model on each sub-tree which are top contributors to the total variance of data, i.e. variance importance factor. Each decision tree has many leaves at the end that contain the most homogeneous data. Each leaf provides the probability of an event of interest in that sub-sample. The average of these leaves across the sub-samples is the mean probability of that event. Many uncorrelated decision trees of sub-samples form a forest or a set of decision trees. These decision trees provide estimates of the parameters or probabilities of that event. The average of these probabilities can be considered better than a single probability estimate.

In the CART model, the splitting of each sample into sub-samples is based on the Gini coefficient. In doing so, the purpose is to divide a sample into two or more homogeneous child sub-samples based on a threshold Gini value or cut-points. The child sub-samples have homogeneous units, whereas they are heterogeneous between themselves. The sample splits until homogeneity or stopping criteria are met in the node(8, 9). The minimum change in heterogeneity in this CART study was set at 0.0001, maintaining a minimum sample of five observations in a leaf. The growth limit was predetermined based on the number of samples in the parent node and child nodes, whereas the maximum depth of the decision tree was fixed at seven for analytical outcomes; it can also be automatically determined if the sample size is very large. Missing values were treated as missing values.

Three distinct strengths of CART make it particularly applicable to analysing complex disease modelling. Firstly, the hierarchical structure of CART models is often more intuitive than traditional regression models because it mimics the heuristics of decision-making with conditional probabilities. Secondly, the CART model can outperform standard regression models when predicting outcomes in nonlinear relationships, and the interaction thresholds may vary with direct and intermediating risk factors. For example, the probability of a diagnosis of CVD in individuals decreases if they have a later onset of cholesterol and increases in the case of an early onset of cholesterol. Thirdly, the CART model affords the data of greater freedom to speak for themselves. On the other hand, regression models are refined by comparing a limited number of covariate specifications.

The CART model performs an exhaustive search over all possible cut-points and predictors. As a result, the precise relationship between a predictor and outcome is not delimited by the inclusion/exclusion of higher-order terms. It is this strength that has seen CART used in a variety of prognostic analyses to identify risk thresholds for killer diseases, such as heart diseases.

References

1. Perianayagam A, Bloom D, Lee J, Parasuraman S, Sekher TV, Mohanty SK, et al. Cohort Profile: The Longitudinal Ageing Study in India (LASI). Int J Epidemiol. 2022;51(4):e167-e76.

2. Bloom DE, Sekher TV, Lee J. Longitudinal Aging Study in India (LASI): new data resources for addressing aging in India. Nat Aging. 2021;1(12):1070-2.

3. WHO. International Statistical Classification of Diseases and Related Health Problems 10th Revision. Geneva: WHO; 2019.

4. Andersen RM. Revisiting the Behavioral Model and Access to Medical Care: Does it Matter? Journal of Health and Social Behavior. 1995;36(1).

5. Andersen R, Newman JF. Societal and Individual Determinants of Medical Care Utilization in the United States. The Milbank Quarterly. 2005;83(4).

6. Pedersen TL. A Tidy API for Graph Manipulation [R package tidygraph version 1.3.1]. In: Pedersen TL, editor. 2024.

7. Therneau T, Atkinson B, Ripley B. Recursive Partitioning and Regression Trees [R package rpart version 4.1.23]. In: Atkinson B, editor. Recursive partitioning for classification, regression and survival trees An implementation of most of the functionality of the 1984 book by Breiman, Friedman, Olshen and Stone2023.

8. Mienye ID, Jere N. A Survey of Decision Trees: Concepts, Algorithms, and Applications. IEEE Access. 2024;12:86716-27.

9. Song YY, Lu Y. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry. 2015;27(2):130-5.