Author/s:
Dipnall, JF;Pasco, JA;Berk, M;Williams, LJ;Dodd, S;Jacka, FN;Meyer, D

Title:
Into the bowels of depression: Unravelling medical symptoms associated with depression by applying machine-learning techniques to a community based population sample

Date:
2016-12-01

Citation:
Dipnall, J. F., Pasco, J. A., Berk, M., Williams, L. J., Dodd, S., Jacka, F. N. & Meyer, D. (2016). Into the bowels of depression: Unravelling medical symptoms associated with depression by applying machine-learning techniques to a community based population sample. Plos One, 11 (12), https://doi.org/10.1371/journal.pone.0167055.

Persistent Link:
https://hdl.handle.net/11343/261030

PLOS | ONE

RESEARCH ARTICLE

Into the Bowels of Depression: Unravelling Medical Symptoms Associated with Depression by Applying Machine-Learning Techniques to a Community Based Population Sample

Joanna F. Dipnall[1,2]*, Julie A. Pasco[1,3,4,5☯], Michael Berk[1,5,6,7,8☯], Lana J. Williams[1,5☯], Seetal Dodd[1,5,6,8☯], Felice N. Jacka[1,6,9,10☯], Denny Meyer[2☯]

1 IMPACT Strategic Research Centre, School of Medicine, Deakin University, Geelong, VIC, Australia, **2** Department of Statistics, Data Science and Epidemiology, Swinburne University of Technology, Melbourne, Victoria, Australia, **3** Melbourne Clinical School-Western Campus, The University of Melbourne, St Albans, VIC, Australia, **4** Department of Epidemiology and Preventive Medicine, Monash University, Melbourne, VIC, Australia, **5** University Hospital Geelong, Geelong VIC Australia, **6** Department of Psychiatry, The University of Melbourne, Parkville, VIC, Australia, **7** Florey Institute of Neuroscience and Mental Health, Parkville, VIC, Australia, **8** Orygen, the National Centre of Excellence in Youth Mental Health, Parkville, VIC, Australia, **9** Centre for Adolescent Health, Murdoch Children's Research Institute, Melbourne, Australia, **10** Black Dog Institute, Sydney, Australia

☯ These authors contributed equally to this work.
* jdipnall@deakin.edu.au

Citation: Dipnall JF, Pasco JA, Berk M, Williams LJ, Dodd S, Jacka FN, et al. (2016) Into the Bowels of Depression: Unravelling Medical Symptoms Associated with Depression by Applying Machine-Learning Techniques to a Community Based Population Sample. PLoS ONE 11(12): e0167055. doi:10.1371/journal.pone.0167055

Editor: Igor Branchi, Istituto Superiore Di Sanita, ITALY

Received: June 6, 2016

Accepted: November 8, 2016

Published: December 9, 2016

Data Availability Statement: The original and cleaned data for the NHANES data used in this study is open access and located at the URL http://wwwn.cdc.gov/Nchs/Nhanes/Search/nhanes09_10.aspx.

Funding: MB is supported by a NHMRC Senior Principal Research Fellowship 1059660. LJW is supported by a NHMRC Career Development Fellowship 1064272. FNJ is supported by an NHMRC Career Development Fellowship 1108125.

Abstract

Background

Depression is commonly comorbid with many other somatic diseases and symptoms. Identification of individuals in clusters with comorbid symptoms may reveal new pathophysiological mechanisms and treatment targets. The aim of this research was to combine machine-learning (ML) algorithms with traditional regression techniques by utilising self-reported medical symptoms to identify and describe clusters of individuals with increased rates of depression from a large cross-sectional community based population epidemiological study.

Methods

A multi-staged methodology utilising ML and traditional statistical techniques was performed using the community based population National Health and Nutrition Examination Study (2009–2010) (N = 3,922). A Self-organised Mapping (SOM) ML algorithm, combined with hierarchical clustering, was performed to create participant clusters based on 68 medical symptoms. Binary logistic regression, controlling for sociodemographic confounders, was used to then identify the key clusters of participants with higher levels of depression (PHQ-9$\geq$10, n = 377). Finally, a Multiple Additive Regression Tree boosted ML algorithm was run to identify the important medical symptoms for each key cluster within 17 broad categories: heart, liver, thyroid, respiratory, diabetes, arthritis, fractures and osteoporosis, skeletal pain,

The author(s) received no specific funding for this work.

Competing Interests: JFD has no conflicts of interest in relation to this manuscript. JAP has recently received grant/research support from the National Health and Medical Research Council (NHMRC), BUPA Foundation, Amgen,/ GlaxoSmithKline/Osteoporosis Australia/Australian and New Zealand Bone and Mineral Society, Western Alliance, Barwon Health, Deakin University and the Geelong Community Foundation. MB has received Grant/Research Support from the NIH, Cooperative Research Centre, Simons Autism Foundation, Cancer Council of Victoria, Stanley Medical Research Foundation, MBF, NHMRC, Beyond Blue, Rotary Health, Geelong Medical Research Foundation, Bristol Myers Squibb, Eli Lilly, Glaxo SmithKline, Meat and Livestock Board, Organon, Novartis, Mayne Pharma, Servier and Woolworths, has been a speaker for Astra Zeneca, Bristol Myers Squibb, Eli Lilly, Glaxo SmithKline, Janssen Cilag, Lundbeck, Merck, Pfizer, Sanofi Synthelabo, Servier, Solvay and Wyeth, and served as a consultant to Astra Zeneca, Bioadvantex, Bristol Myers Squibb, Eli Lilly, Glaxo SmithKline, Janssen Cilag, Lundbeck Merck and Servier. Drs Copolov, MB and Bush are co-inventors of provisional patent 02799377.3-2107-AU02 "Modulation of physiological process and agents useful for same". MB and Laupu are co-authors of provisional patent 2014900627 "Modulation of diseases of the central nervous system and related disorders". MB is supported by a NHMRC Senior Principal Research Fellowship 1059660. LJW is supported by a NHMRC Career Development Fellowship 1064272. SD has received grants/ research support from the Stanley Medical Research Institute, NHMRC, Beyond Blue, ARHRF, Simons Foundation, Geelong Medical Research Foundation, Fondation FondaMental, Eli Lilly, Glaxo SmithKline, Organon, Mayne Pharma and Servier, speaker's fees from Eli Lilly, advisory board fees from Eli Lilly and Novartis, and conference travel support from Servier. FNJ has received Grant/ Research support from the Brain and Behaviour Research Institute, the National Health and Medical Research Council (NHMRC), Australian Rotary Health, the Geelong Medical Research Foundation, the Ian Potter Foundation, Eli Lilly, the Meat and Livestock Board and The University of Melbourne and has received speakers honoraria from Sanofi-Synthelabo, Janssen Cilag, Servier, Pfizer, Health Ed, Network Nutrition, Angelini Farmaceutica, and Eli Lilly. She is supported by an NHMRC Career Development Fellowship (#1108125). DM has received grant/research support from the Australian research Council (ARC), Mental Illness blood pressure, blood transfusion, cholesterol, vision, hearing, psoriasis, weight, bowels and urinary.

Results

Five clusters of participants, based on medical symptoms, were identified to have significantly increased rates of depression compared to the cluster with the lowest rate: odds ratios ranged from 2.24 (95% CI 1.56, 3.24) to 6.33 (95% CI 1.67, 24.02). The ML boosted regression algorithm identified three key medical condition categories as being significantly more common in these clusters: bowel, pain and urinary symptoms. Bowel-related symptoms was found to dominate the relative importance of symptoms within the five key clusters.

Conclusion

This methodology shows promise for the identification of conditions in general populations and supports the current focus on the potential importance of bowel symptoms and the gut in mental health research.

Introduction

Depression is a debilitating illness that is estimated to affect 350 million people globally and is frequently associated with somatic symptoms and other medical conditions [1,2]. The nature and direction of these relationships are often complex, interrelated, and difficult to unravel. Depression classically presents with many and diverse somatic symptoms. The comorbidity of depression with a number of chronic medical conditions, such as Irritable Bowel Syndrome (IBS) [3], ischemic heart disease [4], cancer [5], diabetes [6], osteoporosis [7], thyroid disease [8], and obesity [9], has also been well established. However, these conditions often have bidirectional relationships with depression such that this level of comorbidity and interrelatedness can complicate treatment and stymie efforts to identify causal factors in depression. Thus, the identification of individuals in clusters of comorbid symptoms in depression may reveal new pathophysiological mechanisms and treatment targets.

Due to the complexity and heterogeneity of medical data, previous studies have primarily investigated individual medical conditions linked to depression. The use of "big data" and machine-learning (ML) techniques and algorithms has the ability to handle heterogeneous data without strict constraints and have been demonstrated to unearth key patterns and interactions in health data [10,11]. The mapping of multidimensional data onto two-dimensional maps [12–14] with ML techniques allows the researcher to visualise and interpret the complexity of the data and generate new hypotheses regarding depression.

ML is a vast and expanding field of artificial learning where algorithms improve performance through experiential learning [15]. In the health arena, ML algorithms that learn by training on subsets of data have been used to fit models using supervised ML (i.e. where the objective of the exercise is to establish the main inputs to predict known values) [16], and to find patterns in data using unsupervised ML (i.e. where the objective is to uncover previously unknown patterns and clusters within the data set, without any a priori model defined) [17]. Blending of unsupervised and supervised ML techniques has been used to detect patterns and relationships within large numbers of complex lifestyle-environ variables [18]. Notoriously

Research Fund (MIRF), Victorian Department of Justice, Beyond Blue, Swinburne University of Technology, Federal University. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

Abbreviations: DIPIT, Data Integration Protocol In Ten-Steps; ML, Machine-learning; MART, Multiple Additive Regression Trees; NCHS, National Center for Health Statistics; NHANES, National Health and Nutrition Examination Survey; PHQ-9, Patient Health Questonnaire-9; SOMs, Self-organizing maps.

complex in nature, medical symptom data are ideally suited to blended ML techniques. Utilising the learning properties of ML it is possible to detect, visualize and understand the composition of medical symptoms clusters for those with psychiatric disorders such as depression. [19,20]

ML techniques have been used across a variety of disciplines to explore and model very large quantities of data to discover patterns, unsuspected relationships and useful rules for a specific purpose. Often novel unsuspected and novel interpretations of the data (serendipity) are uncovered. Commercially, these techniques have been used successfully for businesses to learn from their transaction data about the behaviour of their customers, improving their business model by exploiting this knowledge [21]. However, it has only been over the last 10 years that ML techniques have been used in medical research, primarily in neuroscience and biomedicine [22,23]. More recently ML techniques have been used in psychiatry [10], using predominantly very big data sets. Complex survey methodologies are often implemented with population-based data (e.g. oversampling in underrepresented groups, stratification, clustering) and traditional statistical techniques are capable of dealing with this complexity [24]. However, big data techniques on their own do not adequately account for this type of sample. Thus, a blend of both big data ML techniques with traditional statistical techniques has the potential to uncover hidden patterns while accounting for the complex sampling.

The aim of this research was to use data from a large cross-sectional community based population epidemiological study to combine unsupervised and supervised ML algorithms with traditional regression techniques by utilising self-reported medical symptoms to identify and describe clusters of individuals with increased rates of depression from a large cross-sectional community based population epidemiological study.

Methods

Study design and participants

The 2009–2010 National Health and Nutrition Examination Survey (NHANES) (2009–2010) [25] cross-sectional civilian noninstitutionalized population based data were utilised for this study. This study included 18 to 80 year old non-institutionalised US civilians (N≈10,000) and applied a complex four-stage sampling methodology: counties; segments within counties; households within segments; and, individuals within households. Data were collected from 15 locations across 50 US states, with oversampling of subgroups of the population of particular public health interest, to increase the reliability and precision of population estimates [25]. Questionnaire data relating to medical symptoms and demographics were downloaded from the NHANES website and integrated using the Data Integration Protocol In Ten Steps (DIPIT) [26].

Variables were initially selected based on the criterion of relevance to medical symptoms. Analysis was performed to minimise the degree of missing data across the set of medical symptoms. The final set of 68 dichotomous medical symptom variables and an unweighted sample size of 3,922 was used for clustering in this research study. There were 377 participants identified with depression, being representative of the total depressed sample for NHANES during 2009–2010 (i.e. 8% after adjustment for the complex survey sample structure). The imbalanced nature of the data was addressed in this study by identifying clusters with high rates of depression (i.e. high risk clusters) rather than individual participants with depression. This meant that within each high risk cluster the imbalance was much reduced. This was the primary rationale for undertaking the Self-organised Mapping (SOM) and clustering of individuals, thereby allowing the identification of the key clusters significantly associated with depression using binary logistic regression. Finally, the most important medical symptoms for identifying depressed individuals were identified for each of these key clusters.

NHANES received approval from the National Center for Health Statistics (NCHS) research ethics review board and informed consent was obtained from all participants. Use of data from the NHANES 2009–2010 database is approved by the National Center for Health Statistics Research Ethics Review Board (Continuation of Protocol #2005–06).

Study Measurements

A self-reported Patient Health Questonnaire-9 (PHQ-9) [27] was used to assess depressive symptoms ('depression'). This questionnaire consisted of nine items that were summed to form a total score. Those with a total score of 10 or more were considered moderately or severely depressed [28]. The 68 medical symptom data were classified into 17 broad medical categories: heart, liver, thyroid, respiratory, diabetes, arthritis, fractures and osteoporosis, pain (i.e. neck, back, hip pain), blood pressure, cholesterol, vision, hearing, psoriasis, weight, bowels, urine, and if a blood transfusion was received. The self-report demographic and socio-economic variables from the NHANES demographic and questionnaire data components were also utilised [29].

Statistical Methodology

This research implemented two ML algorithms: an unsupervised algorithm, combined with hierarchical clustering, to create the medical symptom clusters and a supervised algorithm to identify and describe the key clusters with a significant relationship with depression. Due to the complex sampling methodology of the NHANES data, traditional binary logistic regression was implemented to identify these key clusters while controlling for potential socio-demographic confounders.

A summary of the statistical methodology, testing regime and results is outlined in Fig 1.

Medical symptom cluster identification

Self-organizing maps (SOMs) were introduced by Kohonen in 1995 [30] as a variant of artificial neural networking, inspired by biological neural networks, and have since been used in many diverse applications across a variety of fields including bioinformatics, engineering, financial analysis, experimental physics, and psychiatry [31,32]. SOMs provide a simple and effective unsupervised ML algorithm for clustering individual participants and visualising high dimensional data in a low dimensional map without any reliance on distributional assumptions.

The SOM identifies clusters by effectively packing the dataset onto a q-dimensional plane where data points "similar" to each other in the original multidimensional data space are then mapped onto nearby areas of the q-dimensional output space. SOMs combine competitive learning with dimensionality reduction by smoothing the clusters with respect to an apriori grid. The SOM is called a topology-preserving map because multi-dimensional input data is represented often by a two dimensional "map" of nodes where topological properties of the input space are maintained.

The steps involved in the SOM competitive ML algorithm involve initially assigning random vector weights to each node (or position on the grid), then randomly choosing data points (participants) from the training data and presenting them to the SOM. The "Best Matching Unit" (BMU) in the map is the node with a vector weight most similar to a data point and nodes within the "neighbourhood" of each BMU are found. With each iteration, the size of this neighbourhood decreases. The vector weights of nodes in the BMU neighbourhood are adjusted closer to their associated data points. The size of these adjustments decrease with each iteration and the magnitude of these adjustments is proportional to the proximity of the

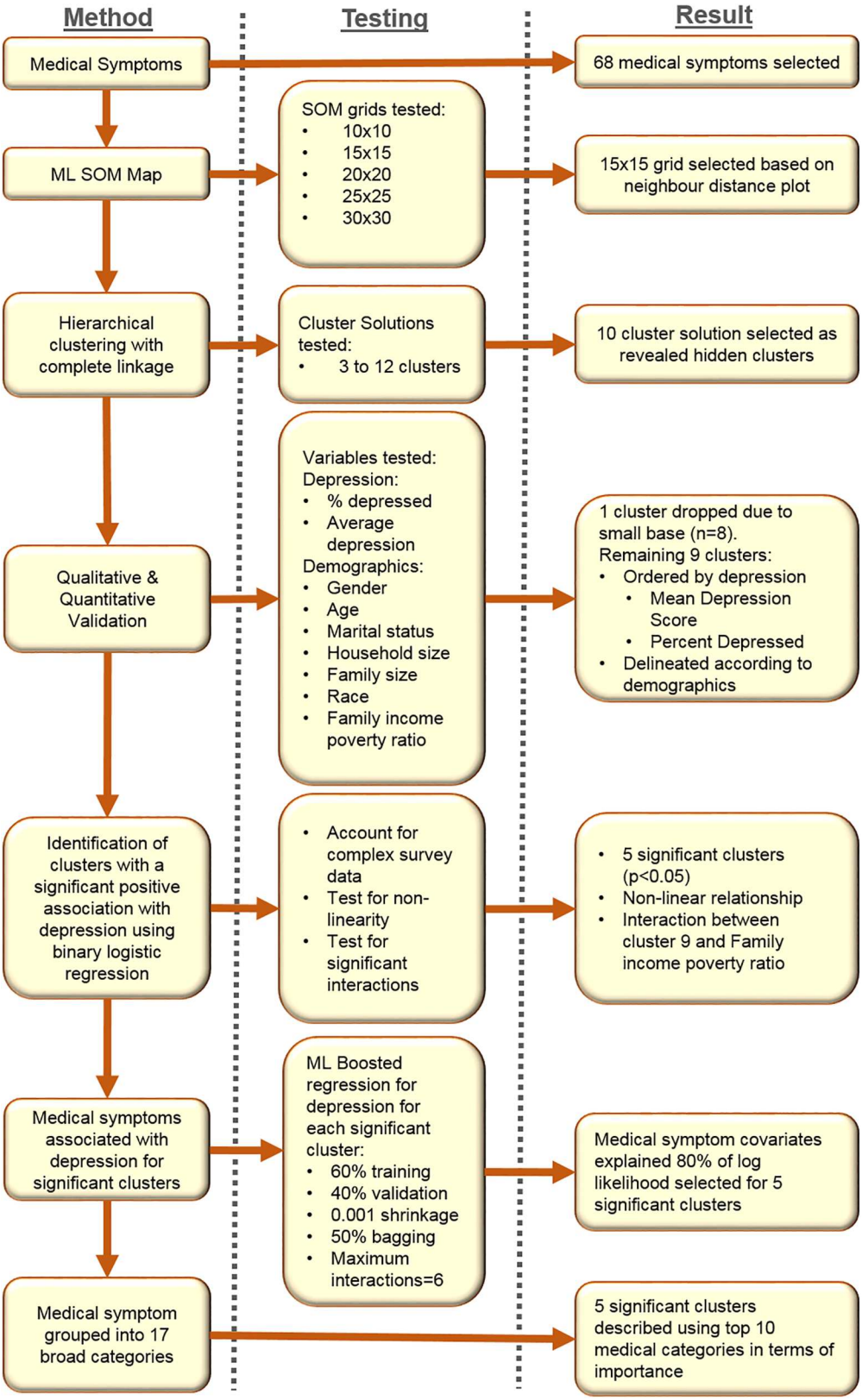


Fig 1. Flowchart of Methods, Testing and Results.

doi:10.1371/journal.pone.0167055.g001

node to the BMU. These steps are repeated for N iterations or until the vector weights for all the nodes converge to their final values.

For this study a hexagonal map topology was used, with five SOM grids tested (10x10, 15x15, 20x20, 25x25, 30x30) to establish a map with suitable nodes. The final solution utilised a 15x15 grid with a learning rate for weight adjustment declining linearly from 5% to 1% over 100 iterations. The unconstrained nature of the SOM technique meant that clusters of nodes form naturally from the medical symptom data on the grid without the influence of the participant's depressive symptom status. Hierarchical clustering, using the complete linkage method [33], was then utilised to group SOM nodes with similar final weights, identifying the final clusters. Three to 12 cluster solutions were considered and the cluster solution with the most differentiation in terms of depression was chosen for further investigation. The clusters were numbered in order of their rates of depression (i.e. frequency and average total PHQ-9 score).

Identification of key clusters with higher depression rates

Quantitative and qualitative investigation, using exploratory statistics of the resultant clusters was used to establish variation with respect to depression rates and demographics.

Demographic factors were included in a binary logistic regression model to identify the key participant clusters with a significant positive relationship with depression, accounting for the complex survey design of NHANES. This model controlled for potential confounders and quantified the probability of depression within each cluster. The cluster with the lowest depression rate was chosen as the reference group. This stage of the analysis was used to identify participant clusters with significant rates of depression in order to identify the important medical symptoms from the ML boosted regression. Only these key clusters were used in the next stage of supervised ML boosted regression. No further investigation was performed on those clusters with non-significant odds ratios for depression.

Medical symptoms most prominent within key clusters

Supervised ML boosted regression [34], translated to a binary logistic regression analysis [35], was implemented for each of the key clusters to identify the most prominent medical symptoms associated with depression within these clusters. This technique has been previously used to identify biomarkers associated with depression [36] and to describe lifestyle clusters associated with depression [18] using data from the NHANES study. Depression was considered as a binary outcome and run for each key cluster using Friedman's Multiple Additive Regression Trees (MART) boosted algorithm [37,38]. Consistent with previous research using this ML algorithm on the 2009 to 2010 NHANES data [36], validation was performed using a random split of each data set into 60% training and 40% validation, a regularization shrinkage parameter of 0.001, with 50% of the residuals used to fit each successive tree (50% bagging) [37]. The maximum number of boosting interactions (i.e. number of terminal nodes plus 1) allowed was six, being marginally higher than the default (i.e. five) and within the recommended range [35]. Whilst this technique has been used for predictive purposes [16], it also has the ability to be used as a variable selection method [36]. This method was used as a variable selection technique to identify the prominent medical symptoms associated with depression within the key clusters [37]. A relative importance (or contribution) of each medical symptom variable for each of the key significant clusters was produced from the ML boosted regression. Higher values of relative importance for a medical symptom within a particular key cluster indicates a stronger relationships with depression in this cluster. This technique for variable reduction has been recognised as effective [39] and previously used to delineate lifestyle clusters associated with depression [18].

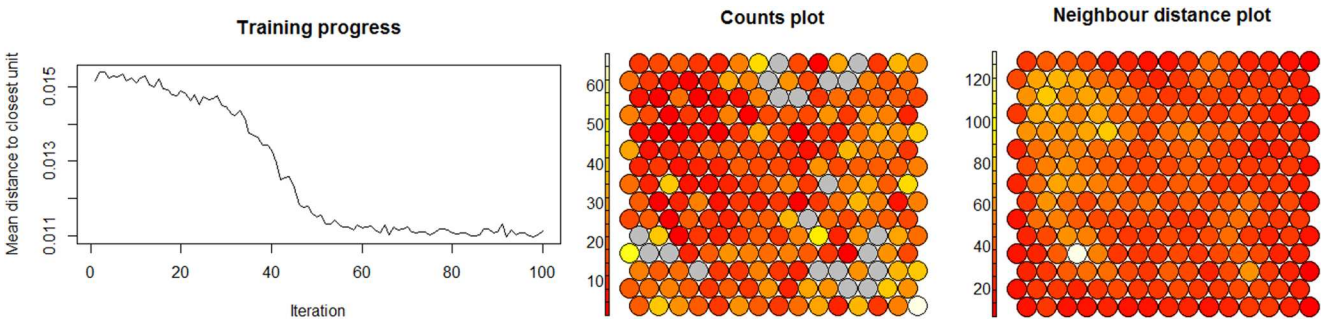


Fig 2. Training progress and SOM plots. Note: The "Training progress" graph indicates as the SOM training iterations distance from each node's weights to the samples represented by that node reduces and plateaus to indicate no more iterations were required. The "Counts plots" indicates reasonable samples were mapped to each node on the map. The "Neighbour distance plot" or U-Matrix indicates the distance between each node and its neighbours.

doi:10.1371/journal.pone.0167055.g002

Those medical symptoms explaining at least 80% of the total log likelihood variation across clusters were used to identify the most important medical symptoms for explaining differences across clusters. Resultant medical symptoms were then grouped into the 17 broad medical categories.

The SOMs and hierarchical clustering were performed in R with the SOM using the Kohonen package [13]. The boosted regression and binary logistic regression statistical procedures were performed using Stata V14 software (StataCorp., 2014), with a Stata plugin for the boosted regression component of the analysis [38].

Results

A summary of the results from the testing is presented in Fig 1.

SOM Clusters

The distance from each node's weights to the sample of people represented by that node was reduced to a minimum plateau as the SOM training iterations progressed, indicating that no more iterations were required (Fig 2). Taking into account the heterogeneous nature of the

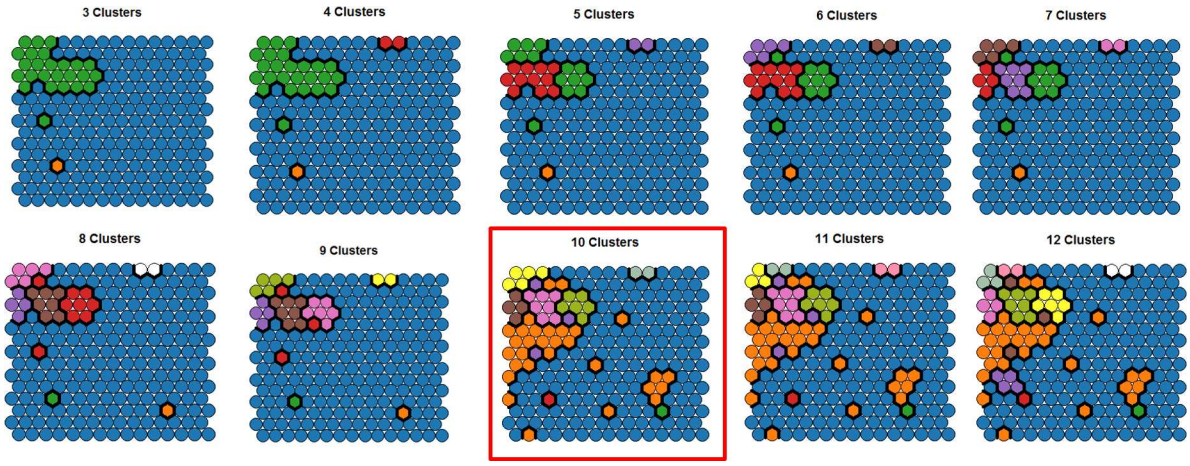


Fig 3. Hierarchical Cluster Options for SOM. Note: Clusters 3 to 12 solutions mapped onto the SOM grid. Colours indicate different clusters. The final 10 cluster solution selected for further analysis has been highlighted with a red border.

doi:10.1371/journal.pone.0167055.g003

Table 1. Frequency Distribution of Initial Depression Ordered SOM Cluster Solution.

Cluster	Frequency	Percent
1	**3,108**	**79.25**
2	34	0.87
3	57	1.45
4	**446**	**11.37**
5	50	1.27
6	83	2.12
7	55	1.4
8	52	1.33
9	29	0.74
10 (Dropped)	8	0.2
Total	**3,922**	**100**

Note: Dominant clusters in bold. Cluster shaded dropped due to very small base (n = 8).

doi:10.1371/journal.pone.0167055.t001

self-reported medical symptom data, the counts plot indicated a reasonable distribution of people numbers across the map. The neighbour distance plot indicated the distances between each node and its neighbours were mostly similar with only a few dissimilar nodes, later identified as outlying clusters (Fig 2).

Three to 12 cluster solutions were considered (Fig 3) and the 10 cluster solution was selected for further investigation because of clear cluster differences in terms of depression rates. There were some isolated nodes in this cluster solution, later confirmed as outliers.

The final 10 cluster solution contained two dominant clusters (Table 1). One cluster was dropped from further analysis due to very low frequency (n = 8), leaving 9 of the 10 clusters for further analysis.

Cluster validation

Initial investigation into the relationship between the remaining nine participant clusters and the depression measures revealed that the clusters exhibited an order with respect to both the percentage of participants depressed within each cluster and the average depression score (Fig 4).

An initial inspection of the socio-demographics for the nine clusters (Table 2) showed clear differences. Due to the small frequencies for many of the clusters, only a qualitative investigation of socio-demographic differences was performed. Cluster 1 (n = 3,108) exhibited socio-demographics closest to the total across all cluster participants. Cluster 2 (n = 34) consisted of mostly male, non-Hispanic white with a high family income poverty ratio [40,41] and who were less likely to have never married. Cluster 3 (n = 57) consisted mostly of male, non-Hispanic white, older, married / with a partner, a household size of around two people, and a low family income poverty ratio. Cluster 4 (n = 446) members were more likely to be female, non-Hispanic white, middle aged, with a low family income poverty ratio, and less likely to have never been married. Cluster 5 (n = 50) were more likely to be older, non-Hispanic black, with a low family income poverty ratio, and less likely to never have been married. Cluster 6 (n = 83) members were more likely to be male, older, less than three members in the household, non-Hispanic white, and of low to mid family income poverty ratio, and less likely to have never been married. Cluster 7 (n = 55) were more likely to be middle aged, Mexican / Hispanic, with a low family income poverty ratio and less likely to have never been married. Cluster 8 (n = 52) were more likely to be female, older, non-Hispanic white, around two members in the household, with low family income poverty ratio and less likely to have been

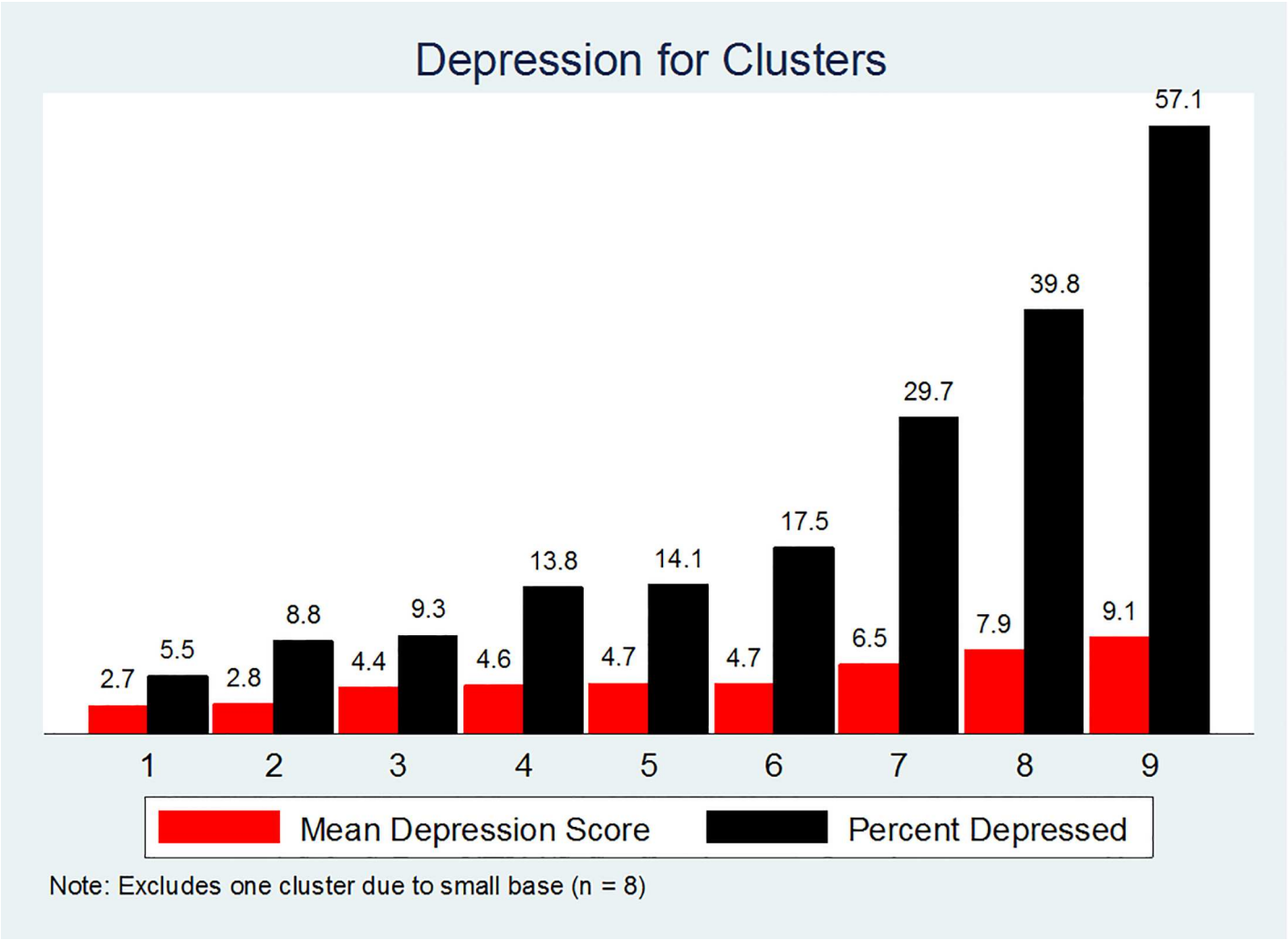


Fig 4. Mean depression scores and percent depression across final depression clusters. Note: "Mean Depression Score" is the average total PHQ-9 score which ranged from 0 to 27. "Percent Depressed" based on a total PHQ-9 $\geq$ 10.

doi:10.1371/journal.pone.0167055.g004

married. Cluster 9 (n = 29) were more likely to be young Mexican / Hispanic, with a large household and low family income poverty ratio.

Identification of key clusters with higher depression rates

The final binary logistic regression with depression as the outcome took into account the complex survey data of NHANES, as well as non-linearity, interactions and potential confounders (Table 3). The test for goodness of fit were not significant for the model indicating a good fit to the data ($F(9,8) = 1.77$, $p = 0.216$) [42]. Clusters 4 and 6 to 9 had significantly higher rates of depression than cluster 1 after controlling for the potential socio-demographic confounders. These five clusters were considered the key clusters for further analysis. Since the odds ratios for depression for clusters 2, 3 did not significantly differ from cluster 1 these clusters were excluded from future analysis. A significant interaction was found between the cluster with the highest rate of depression (cluster 9) and the family income poverty ratio ($p = 0.036$) (Fig 5). Thus, the relationship between the probability of depression and cluster 9 varied depending upon the rate of the family poverty income ratio.

Table 2. Demographic Profile Across SOM Clusters.

CLUSTER	Total	1	2	3	4	5	6	7	8	9
Sample (n)*	3,914	3,108	34	57	446	50	83	55	52	29
Demographics										
Gender:										
Male	49.6%	50.8%	60.4%	60.2%	38.9%	43.1%	63.9%	41.2%	31.2%	57.4%
Female	50.4%	49.2%	39.6%	39.8%	61.1%	56.9%	36.1%	58.8%	68.8%	42.6%
Mean age (years)	42.44	42.08	48.17	53.05	46.09	51.60	56.24	44.30	55.93	38.67
Marital status:										
Never	19.3%	20.4%	15.0%	11.6%	15.8%	14.6%	7.8%	14.7%	7.6%	19.6%
Married/Partner	65.3%	65.5%	66.4%	76.6%	64.7%	65.2%	63.9%	66.8%	48.9%	63.6%
Widowed/Divorced/Separated	15.4%	14.1%	18.6%	11.8%	19.6%	20.3%	28.4%	18.6%	43.5%	16.8%
Mean household size	3.22	3.23	3.19	2.23	3.07	2.95	2.64	3.26	2.55	3.94
Mean family size	3.02	3.03	3.07	2.94	2.91	2.75	2.49	2.94	2.39	3.88
Race:										
Mexican/Hispanic	14.3%	14.4%	6.0%	17.3%	15.3%	13.3%	5.6%	27.0%	2.6%	45.5%
Non-Hispanic white	67.5%	67.6%	78.4%	66.3%	67.3%	54.0%	74.2%	55.4%	81.3%	26.0%
Non-Hispanic black	11.4%	10.9%	15.6%	5.9%	12.8%	27.0%	17.4%	14.6%	10.1%	16.5%
Other	6.7%	7.1%	0.0%	10.4%	4.6%	5.7%	2.8%	2.9%	6.1%	12.0%
Family income poverty ratio:**										
Low	31.0%	29.2%	29.9%	39.8%	35.7%	42.6%	32.7%	64.9%	58.9%	68.8%
Middle	24.0%	24.5%	10.5%	19.0%	22.9%	23.4%	31.3%	15.9%	21.5%	11.2%
High	45.0%	46.4%	59.6%	41.2%	41.5%	33.9%	36.0%	19.2%	19.6%	20.0%
Mean family income poverty ratio:										
(Note: 1 = poverty line)	3.03	3.16	3.39	2.80	2.89	2.68	2.91	1.82	2.09	1.68

Note: Figures quoted take account of the survey design of NHANES with 15 strata, 31Primary Sampling Units (PSU).

*Total sample size varies per demographic as base includes all those with a depression score and valid answer given for demographic.

**Family income poverty ratio represents the ratio of family or unrelated individual income to their appropriate poverty threshold where groupings are based on eligibility for Special Supplemental Nutrition Program for Women, Infants, and Children (WIC): Low = 0.00–1.85 family income poverty ratio, Middle = >1.85–3.50 family income poverty ratio, and High = >3.50 and above family income poverty ratio.

doi:10.1371/journal.pone.0167055.t002

Medical symptoms most prominent within key clusters

ML boosted regression was used to establish which medical symptoms were associated with depression for each of the five key significant clusters. The top medical symptom variables explaining approximately 80% of the total log likelihood for each cluster were selected for categorisation and further investigation. Bowel symptoms (e.g. bowel movements per week, stool type) dominated the relative importance percentage across all the five key clusters (Fig 6). Further investigation into the top 3 to 10 ranked medical categories from the ML boosted regression found that bowel, pain and urine symptoms consistently exhibiting a relatively high importance percentage for each of the key clusters.

The top 10 key medical symptom categories for the five key significant clusters indicated that each cluster exhibited different medical symptoms (Fig 7). However, bowel symptoms were consistently included in the highest ranked medical symptoms across all five significant depressive key clusters. In addition, the bowel symptoms dominated for cluster 7 and cluster 9, and had relatively high importance (i.e. >5%) for four of the five key clusters. Pain symptoms had the highest relative importance in cluster 4 and urine symptoms had relatively high importance (i.e. <10%) for two of the five key clusters. Whilst hearing symptoms were important in all five of the key clusters, they only dominated in cluster 8.

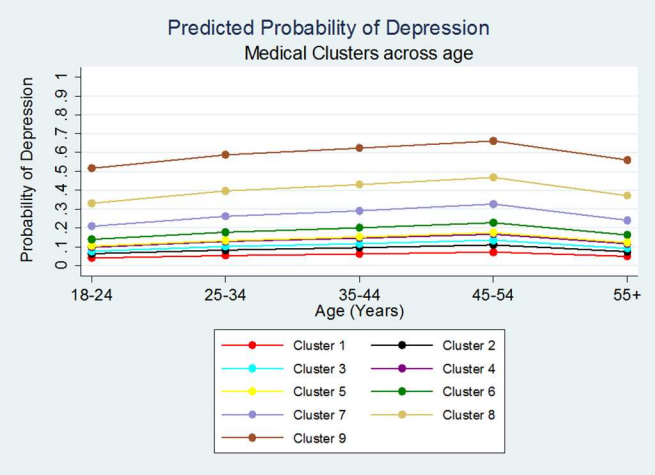
Table 3. Binary Logistic Regression Model Odds Ratios with 95% Confidence Intervals.

Depression	OR	p-value	95% CI Low	95% CI High
Cluster 1 (reference)	1.00			
Cluster 2	1.67	0.341	0.55	5.04
Cluster 3	1.98	0.151	0.76	5.20
Cluster 4	2.24	**<0.001**	1.56	3.24
Cluster 5	2.10	0.180	0.68	6.43
Cluster 6	3.78	**<0.001**	2.17	6.57
Cluster 7	4.61	**<0.001**	2.21	9.63
Cluster 8	7.80	**0.001**	2.86	21.33
Cluster 9	6.33	**0.010**	1.67	24.02
Cluster 9 X Family income poverty ratio	2.00	**0.036**	1.05	3.81
Gender				
Male (reference)	1.00			
Female	1.86	**0.002**	1.31	2.64
Age group				
18–24 years (reference)	1.00			
25–34	1.37	0.326	0.71	2.63
35–44	1.61	0.177	0.79	3.29
45–54	1.92	**0.023**	1.11	3.34
55+	1.22	0.545	0.62	2.39
Marital status				
Never married (reference)	1.00			
Married/living with partner	0.54	**0.007**	0.35	0.82
Widowed/Divorced/Separated	0.79	0.172	0.55	1.12
Gender				
Race				
Non-Hispanic white (reference)	1.00			
Mexican American / Hispanic	0.88	0.368	0.67	1.17
Non-Hispanic Black	1.17	0.436	0.77	1.76
Other	0.77	0.391	0.42	1.43
Education				
Grades 11 and below (reference)	1.00			
High School / GED Equivalent	0.43	**0.039**	0.20	0.95
Some College / AA / College or Above	0.59	**0.008**	0.40	0.85
Family income poverty ratio	0.60	**0.002**	0.45	0.80
Education X Family income poverty ratio				
Grades 11 and below (reference)	1.00			
High School / GED Equivalent	1.43	0.073	0.96	2.11
Some College / AA / College or Above	1.24	0.089	0.96	1.58
Constant	0.16	<0.001	0.08	0.34

Note: OR = Odds Ratio, CI = Confidence Interval. Multivariate logistic model taking account of complex survey methodology (N = 3,584, 15 Strata, 32 PSUs). Bold p-values indicate significant p<0·05. Cluster 9 OR = 12.67 (95% CI: 1.75, 91.56) taking into account the interaction.

doi:10.1371/journal.pone.0167055.t003

The individual clusters showed clear delineation with respect to medical conditions. The top three medical symptoms for cluster 4 related to the skeletal symptoms of pain, fractures and osetoporosis, and bowel symptoms. Cluster 6 was dominated by urinary medical symptoms. Cluster 7 was clearly dominated by bowel medical symptoms. Cluster 8 was a generally



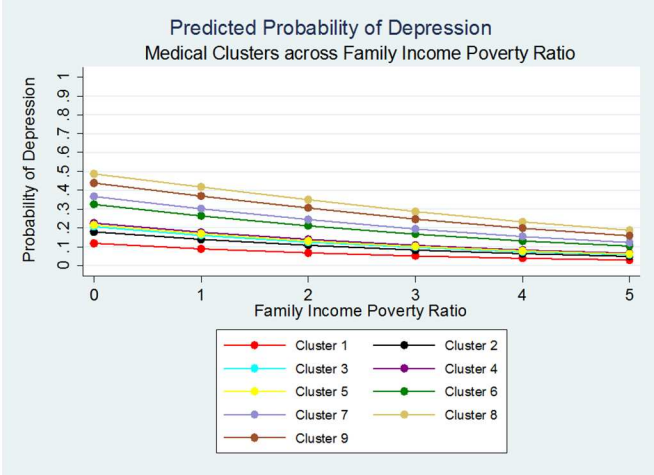
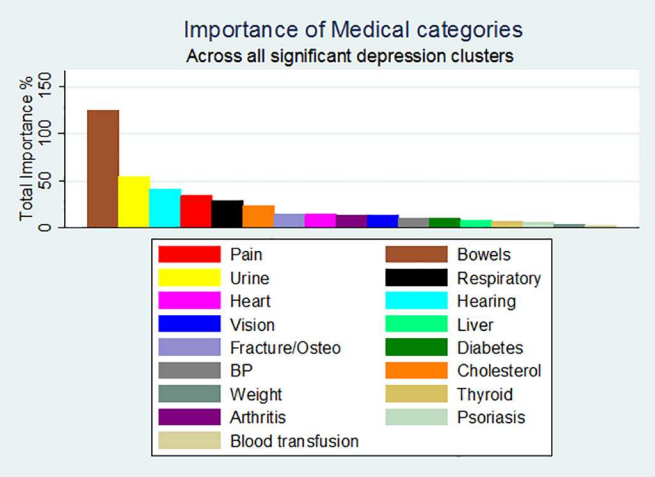
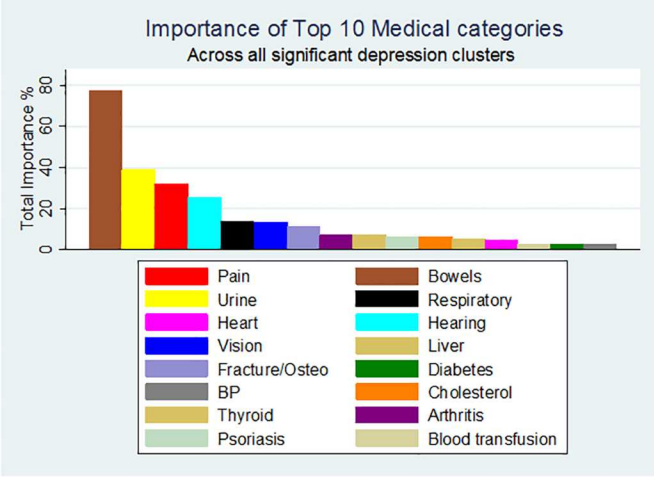
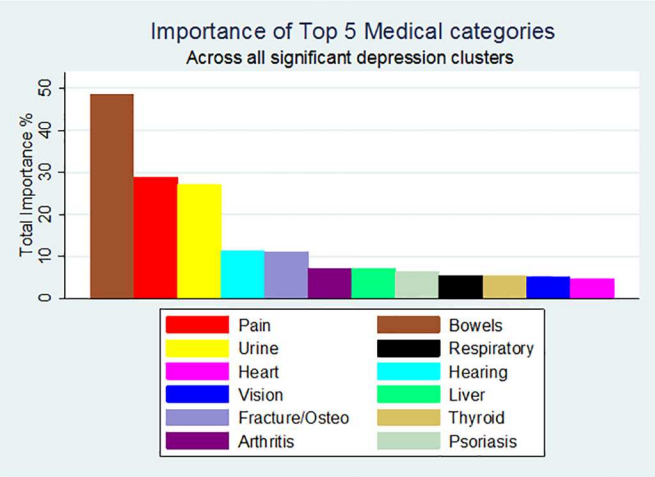


Fig 5. Predicted probability of depression across age and family income poverty ratio for each cluster.

doi:10.1371/journal.pone.0167055.g005







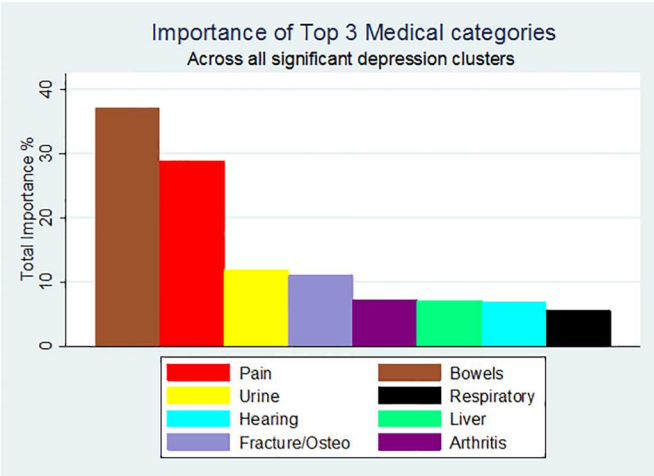


Fig 6. Importance of medical categories that make up the key significant clusters. Note: Based on total boosted relative importance percentage across all clusters. Summed percentage from boosted regression across all five key significant clusters, thus total >100%.

doi:10.1371/journal.pone.0167055.g006

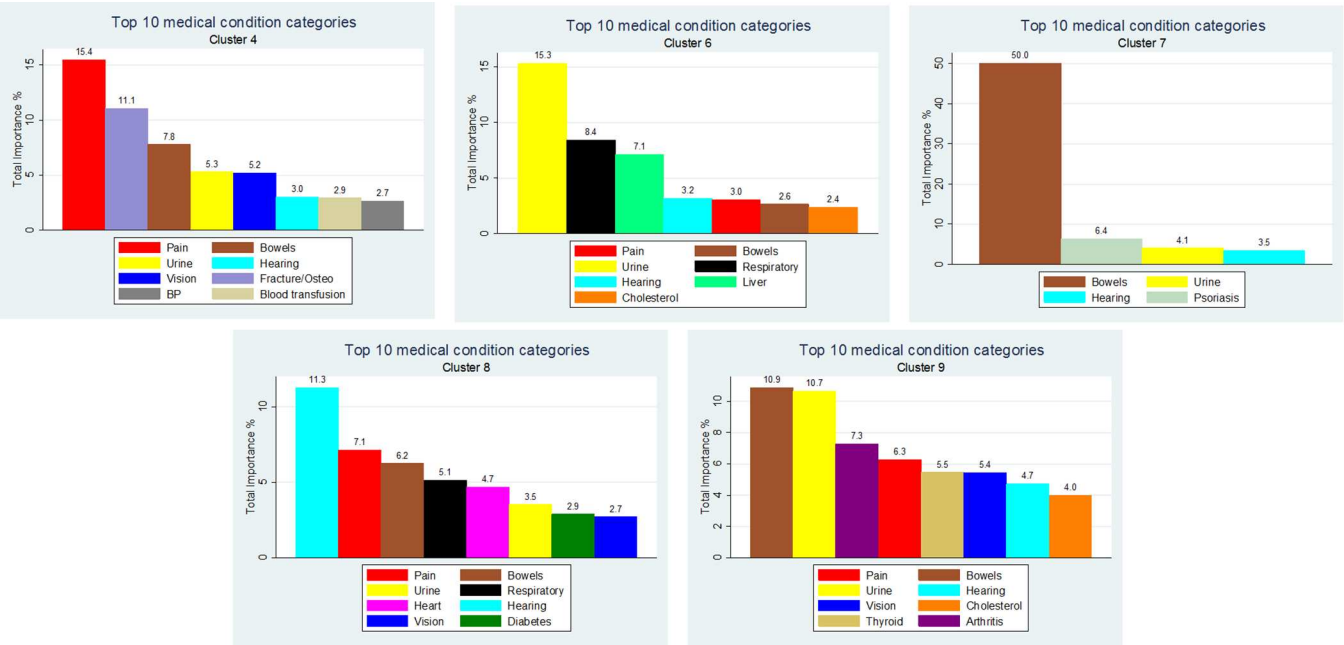


Fig 7. Total percentage importance of medical conditions for each key significant cluster. Note: Clusters presented in order of percent depressed. Note: Percentage sum does not take account of direction of relationship.

doi:10.1371/journal.pone.0167055.g007

unwell cluster with the top five medical symptoms related to hearing, pain, bowels, respiratory and heart. Finally the top two medical symptoms for cluster 9 related to bowels and urine.

Discussion

Irrespective of country, research has consistently found a high level of comorbidity between specific (e.g. sleep, appetite) and nonspecific symptoms and depression [43,44] but it has been difficult to identify the key somatic symptoms most prominent in this condition. This study utilized two machine learning techniques, complemented by traditional binary logistic regression analyses, to detect complex interactions between large numbers of medical symptoms in order to identify those most strongly linked to depression in an atheoretical manner. ML techniques have been used in the area of big data informatics in mental health. For example, text analysis [45] and regression models [46] have been used to predict the risk of suicide from clinical notes, but these techniques have not previously been used to investigate the relationship between depression and medical symptoms using epidemiological community based population data. The visual simplification of complex medical symptom data into clusters, using SOM, allows the researcher to easily identify the strength of the similarities across the map. The ML SOM's intention to mimic an artificial network that learns, without supervision, has proven effective in creating nodes, subsequently grouped into clusters identified by a standard hierarchical clustering. Nine clusters of participants based on medical symptoms were found using the unsupervised graphical SOM ML technique. Traditional binary logistic regression showed that five of the nine clusters were characterised by higher rates of depression after controlling for potential confounders and taking account of the complex survey methodology of the population data.

A boosted regression ML algorithm was used to provide a relative importance percentage for each medical symptom for each of the five key significant clusters, allowing the easy grouping of symptoms into medical categories. The ML boosted regression algorithm was able to

untangle the array of medical symptoms and detect three key medical condition categories as being particularly related to depression: bowel, pain and urinary symptoms. Of these categories, bowel symptoms dominated, validating previous research regarding the high comorbidity between gut symptoms and IBS with common mental disorders, including depression [3,47].

Gut disorders in particular share links with depression. Irritable bowel syndrome (IBS) [3] has been found to be closely associated with mental health conditions. IBS is not only comorbid with psychiatric conditions, but also comorbid with non-gastrointestinal somatic disorders [48]. Crohn's disease [49] and gastro-oesophageal reflux disease (GORD) [50] are similarly associated with higher rates of mood disorders than would be expected by chance. All these interrelationships impact on the quality of life, treatment compliance, length of stay in hospitals, costs of health care, morbidity and possibly mortality of individuals affected.

Medical symptoms relating to stool type and frequency and constipation were included in the bowel categorisation for this study, and these indicators have all been related to mood [51]. Recently, ML boosted regression has identified an association between the gastrointestinal biomarker of bilirubin with depression [36] and bilirubin has been linked to varying stool type based on the speed at which the intestinal contents travel through the bowel [52].

There is an increasing focus in medical research on the role of symbiotic gut microbiota in health and disease, including mental health. Indeed, the human gut microbiota, and what is termed the 'gut-brain axis', are now increasingly regarded as potentially critical drivers of mood and behaviour, with much of the biological dysregulation associated with depressive symptoms and the diagnosis of clinical depression influenced by the gut microbiota [53]. Such microbiota-influenced dysregulation involves inflammatory, metabolic, oxidative stress, HPA axis, neurotransmitter/neuropeptide, brain plasticity and other systems [54]. Moreover, the normal intestinal barrier function is compromised in depression [55]. This 'leaky gut' allows intestinal-microbe-derived lipopolysaccharide (LPS), an endotoxin, to gain access to the periphery. Even very low levels of LPS can provoke much of the aforementioned biological dysregulation noted in depression.

Importantly, many of the lifestyle and environmental factors connected to depression have a detrimental influence on the composition of the normal human microbiota. As just one example, unhealthy dietary patterns that increase the risk for depression [56] also diminish microbial diversity [57]. Long-term, habitual diets are one of the strongest influences on gut microbial composition, determining an individual "enterotype" [58], however dietary change can prompt change in gut microbiota composition within 24 hours [59]. The consumption of complex carbohydrates, plant-based foods/fruits and vegetables [58,60] positively influences microbial composition, synthesis of anti-inflammatory short chain fatty acids, and host health. Conversely, high fat diets trigger microbial dysbiosis, intestinal permeability ('leaky gut') and inflammation [61]. We have previously demonstrated that healthy dietary patterns are associated with a reduced likelihood of depressive symptoms in adults participating in the NHANES [62]. This suggests that unhealthy dietary behaviors may be a key factor negatively influencing both gut health and depression, with bowel symptoms signifying poor gut health.

Strengths and Limitations

The strengths of this study lie in the benefits of using both unsupervised and supervised ML techniques to identify patterns in data, using a large number of heterogeneous self-reported medical symptoms to form five clusters of individuals with relatively high rates of depression, most likely to have remained hidden using traditional statistical techniques. The largest cluster of participants (cluster 4, n = 446) comprised 7% moderately and 7% severely depressed participants; this compares to rates of 5% and 3% respectively in the general 2009 to 2010 US

population in NHANES. The remaining key clusters (6 to 9) consisted of smaller groups of participants, with 15% moderately and 14% severely depressed participants overall. A main limitation with this study is the cross-sectional nature of the NHANES data that restricts the ability to infer causality. However, the use of this community population based survey data has the advantage of being representative of the large US population sampled during 2009 to 2010. The large number of participants included in this study, with its rigorous complex survey sampling methodology, ensures the data possess a good description of the relative characteristics of the civilian noninstitutionalised US population. As compared to other methods of data gathering, surveys are able to extract data that closely mirror attributes of the larger population.

It is acknowledged that the PHQ-9 instrument relates to depressive symptoms, and does not represent a clinical diagnosis of depression. Thus, this self-report instrument may have missed less severe cases of depression [27,28] exaggerating the imbalance in the data. Furthermore, the depression symptoms picked up by the PHQ-9 instrument for this study, such as fatigue, psychomotor problems, or insomnia are symptoms very common in medical conditions. Thus, it was not surprising that the results from this study confirmed prior research identifying depressive symptoms being often elevated in people with medical symptoms [63]. The relationship between medical symptoms and depression is complex and often bidirectional. However, the identification of the dominant medical symptoms, such as those of the bowel cluster in this study, may be used to improve screening tools for depression in medically ill patients and to shed light on possible pathogenic processes. It is acknowledged that individuals with depression are more likely to report somatic conditions, and IBS has been found to be a disorder with a psychosomatic aspect [47]. However, the NHANES study is considered representative of the US noninstitutionalised civilian population and has been used to produce health statistics for the US and in many studies investigating depression (e.g. to examine the prevalence, treatment and control of depressive symptoms [64]).

We addressed the limitation of the imbalance in the data of having only approximately 8% of the sample classified with depression by including only those clusters with high depression rates, hence reducing the impact of this imbalance on our analysis.

There are potential limitations in using the proposed ML techniques. The SOM can become conceptually expensive as the number of variables and the grid size increases, causing the number of distances the algorithm needs to compute to increase exponentially. In addition, the SOM requires a value for each variable for each participant in order to generate a map, so missing data poses issues for map generation with SOMs. Alternative less computer intensive traditional statistical techniques, such as k-means clustering or latent class analysis, could have been used. However, the SOM algorithm has been found to provide better results than either of these methods in the case of large data sets [65–67] such as used in this study.

The ML boosted regression has the advantage of automatically incorporating interaction effects when evaluating variable importance which is not possible with traditional statistical regression modelling [37]. Also, variable selection processes, such as stepwise or regularized regression make variable selection difficult when there are highly correlated predictors as is the case with medical symptoms. The boosted regression overcomes this problem by reducing the number of selected variables at each iteration thereby being able to deal with highly correlated variables. However, ML boosted regression can fail to perform well with small data sets [68]. In addition, the training process can be computationally memory intensive due to the fact that trees are built sequentially, requiring advanced computing capability such as parallel processing. In addition, the regularization implemented to reduce the effects of overfitting can mean the optimal number of iterations for a suitable shrinkage parameter can be considerably large [69].

Whilst this study performed validation using a random split of data into 60% training and 40% validation at the ML boosted regression stage, no validation of the methodology was performed on a separate data set using self-reported medical symptom data. However, this methodology has been successfully implemented to identify lifestyle clusters associated with depression [70].

Conclusion

This study implemented two ML algorithms and a standard binary logistic regression to identify and describe clusters of individuals with higher rates of depression based on self-reported medical symptoms in a large, cross-sectional epidemiological community based population study. Bowel symptoms, covering bowel frequency and stool type, were identified as the predominant concurrent symptom category for the key clusters with a significant positive relationship with depression across 17 varied medical symptom categories. This study encourages the future use of machine learning techniques to compliment traditional statistical approaches in the analysis of epidemiological studies to assist clinicians detect potential latent associations that can be further refined and clarified. This study also supports a research focus on the potential importance of the bowel symptoms, the gut and its resident microbiota in mental health research.

Acknowledgments

MB is supported by a NHMRC Senior Principal Research Fellowship 1059660.

LJW is supported by a NHMRC Career Development Fellowship (GNT1064272).

FNJ is supported by an NHMRC Career Development Fellowship 1108125.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

The authors would like to thank the referees of this issue for their valuable comments and suggestions that have improved this paper.

Author Contributions

Conceptualization: JFD.

Formal analysis: JFD.

Methodology: JFD DM.

Software: JFD.

Visualization: JFD.

Writing – original draft: JFD.

Writing – review & editing: JFD JAP MB LJW SD FNJ DM.

References

1. Sanna L, Stuart AL, Pasco JA, Kotowicz MA, Berk M, et al. (2013) Physical comorbidities in men with mood and anxiety disorders: a population-based study. BMC Med 11: 1.
2. Sanna L, Stuart AL, Pasco JA, Jacka FN, Berk M, et al. (2014) Atopic disorders and depression: findings from a large, population-based study. J Affect Disord 155: 261–265. doi: 10.1016/j.jad.2013.11.009 PMID: 24308896
3. Fond G, Loundou A, Hamdani N, Boukouaci W, Dargel A, et al. (2014) Anxiety and depression comorbidities in irritable bowel syndrome (IBS): a systematic review and meta-analysis. Eur Arch Psychiatry Clin Neurosci 264: 651–660. doi: 10.1007/s00406-014-0502-z PMID: 24705634

4. Kronish IM, Carson AP, Davidson KW, Muntner P, Safford MM (2012) Depressive symptoms and cardiovascular health by the american heart association's definition in the reasons for geographic and racial differences in stroke (REGARDS) study. PLoS One 7: e52771. doi: 10.1371/journal.pone.0052771 PMID: 23300767
5. Massie MJ (2004) Prevalence of depression in patients with cancer. Monographs-National Cancer Institute 32: 57–71.
6. Mezuk B, Eaton WW, Albrecht S, Golden SH (2008) Depression and type 2 diabetes over the lifespan a meta-analysis. Diabetes Care 31: 2383–2390. doi: 10.2337/dc08-0985 PMID: 19033418
7. Fernandes BS, Hodge JM, Pasco JA, Berk M, Williams LJ (2016) Effects of depression and serotonergic antidepressants on bone: mechanisms and implications for the treatment of depression. Drugs Aging 33: 21–25. doi: 10.1007/s40266-015-0323-4 PMID: 26547857
8. Harris B, Othman S, Davies J, Weppner G, Richards C, et al. (1992) Association between postpartum thyroid dysfunction and thyroid antibodies and depression. Bmj 305: 152–156. PMID: 1515829
9. Luppino FS, de Wit LM, Bouvy PF, Stijnen T, Cuijpers P, et al. (2010) Overweight, obesity, and depression: a systematic review and meta-analysis of longitudinal studies. Arch Gen Psychiatry 67: 220–229. doi: 10.1001/archgenpsychiatry.2010.2 PMID: 20194822
10. Passos IC, Mwangi B, Kapczinski F (2016) Big data analytics and machine learning: 2015 and beyond. The Lancet Psychiatry 3: 13–15. doi: 10.1016/S2215-0366(15)00549-0 PMID: 26772057
11. Monteith S, Glenn T, Geddes J, Bauer M (2015) Big data are coming to psychiatry: a general introduction. International journal of bipolar disorders 3: 1–11.
12. Kohenen T (1997) Self-Organizing Maps, Vol. 30 of Lecture Notes in Information Sciences. Springer.
13. Wehrens R, Buydens LM (2007) Self-and super-organizing maps in R: the Kohonen package. J Stat Softw 21: 1–19.
14. Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biol Cybern 43: 59–69.
15. Mitchell TM (1997) Machine learning. 1997. Burr Ridge, IL: McGraw Hill 45.
16. Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, et al. (2016) Cross-trial prediction of treatment outcome in depression: a machine learning approach. The Lancet Psychiatry.
17. Arnrich B, Setz C, La Marca R, Tröster G, Ehlert U (2010) Self Organizing Maps for Affective State Detection. Machine Learning for Assistive Technologies: 45.
18. Joanna F Dipnall JAP, Michael Berk, Lana J Williams, Seetal Dodd,Felice, N Jacka DM (2016) Why so GLUMM? Detecting depression clusters through Graphing Lifestyleenvirons Using Machine-learning Methods (GLUMM). Eur Psychiatry.
19. Vesanto J, Alhoniemi E (2000) Clustering of the self-organizing map. Neural Networks, IEEE Transactions on 11: 586–600.
20. Van Hulle MM (2012) Self-organizing maps. Handbook of Natural Computing: Springer. pp. 585–622.
21. Linoff GS, Berry MJ (2011) Data Mining Techniques: For Marketing, Sales, And Customer Relationship Management Author: Gordon S. Linoff, Michael J. Be.
22. Chaovalitwongse W, Pardalos PM, Xanthopoulos P (2010) Computational Neuroscience: Springer.
23. Seref O, Kundakcioglu OE, Pardalos PM (2007) Data mining, systems analysis, and optimization in biomedicine: American Institute of Physics Inc.
24. Lumley T (2004) Analysis of complex survey samples. Journal of Statistical Software 9: 1–19.
25. Centers for Disease Control and Prevention National Center for Health Statistics (2013) National Health and Nutrition Examination Survey: Analytic Guidelines, 1999–2010 U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
26. Dipnall JF, Berk M, Jacka FN, Williams LJ, Dodd S, et al. (2014) Data Integration Protocol In Ten-steps (DIPIT): A new standard for medical researchers. Methods.
27. Kroenke K, Spitzer RL (2002) The PHQ-9: a new depression diagnostic and severity measure. Psychiatric Annals 32: 509–515.
28. Kroenke K, Spitzer RL, Williams JB (2001) The PHQ-9. J Gen Intern Med 16: 606–613. doi: 10.1046/j.1525-1497.2001.016009606.x PMID: 11556941
29. (CDC). CfDCaP (2009–2010) National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Questionnaire.: Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.
30. Kohonen T (1995) Self-Organizng Maps-Springer Series in Information Sciences, vol. 30. Berlin: Springer Verlag.

31. Gabor A, Leach R, Dowla F (1996) Automated seizure detection using a self-organizing neural network. Electroencephalogr Clin Neurophysiol 99: 257–266. PMID: 8862115

32. Magdolen J, Rappelsberger P, Dorffner G, Flexer A, Winterer G (1997) Evaluating multi-layer perceptrons and self-organising feature maps as a tool for identifying psychiatric disorders in EEG. Psychiatry Research: Neuroimaging 68: 171–172.

33. Köhn HF, Hubert LJ (2006) Hierarchical cluster analysis. Wiley StatsRef: Statistics Reference Online.

34. Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences 55: 119–139.

35. Hastie T, Tibshirani R, Friedman J, Franklin J (2005) The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer 27: 83–85.

36. Dipnall JF, Pasco JA, Berk M, Williams LJ, Dodd S, et al. (2016) Fusing Data Mining, Machine Learning and Traditional Statistics to Detect Biomarkers Associated with Depression. PLoS One 11: e0148195. doi: 10.1371/journal.pone.0148195 PMID: 26848571

37. Friedman J, Hastie T, Tibshirani R (2001) The elements of statistical learning: Springer series in statistics Springer, Berlin.

38. Schonlau M (2005) Boosted regression (boosting): An introductory tutorial and a Stata plugin. Stata Journal 5: 330.

39. Friedman J, Hastie T, Tibshirani R (2000) Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics 28: 337–407.

40. Black MM, Cutts DB, Frank DA, Geppert J, Skalicky A, et al. (2004) Special Supplemental Nutrition Program for Women, Infants, and Children participation and infants' growth and health: a multisite surveillance study. Pediatrics 114: 169–176. PMID: 15231924

41. Bureau UC (2008) Current Population Survey: Definitions and explanations. Population Division, Fertility & Family Statistics Branch.

42. Archer KJ, Lemeshow S (2006) Goodness-of-fit test for a logistic regression model fitted using survey sample data. Stata Journal 6: 97–105.

43. Simon GE, VonKorff M, Piccinelli M, Fullerton C, Ormel J (1999) An international study of the relation between somatic symptoms and depression. N Engl J Med 341: 1329–1335. doi: 10.1056/NEJM199910283411801 PMID: 10536124

44. Kapfhammer H (2006) Somatic symptoms of depression. Dialogues Clin Neurosci 8: 227. PMID: 16889108

45. Poulin C, Shiner B, Thompson P, Vepstas L, Young-Xu Y, et al. (2014) Predicting the risk of suicide by analyzing the text of clinical notes. PLoS One 9: e85733. doi: 10.1371/journal.pone.0085733 PMID: 24489669

46. Tran T, Phung D, Luo W, Venkatesh S (2015) Stabilized sparse ordinal regression for medical risk stratification. Knowledge and Information Systems 43: 555–582.

47. Mykletun A, Jacka F, Williams L, Pasco J, Henry M, et al. (2010) Prevalence of mood and anxiety disorder in self reported irritable bowel syndrome (IBS). An epidemiological population based study of women. BMC Gastroenterol 10: 1.

48. Whitehead WE, Palsson O, Jones KR (2002) Systematic review of the comorbidity of irritable bowel syndrome with other disorders: what are the causes and implications? Gastroenterology 122: 1140–1156. PMID: 11910364

49. Persoons P, Vermeire S, Demyttenaere K, Fischler B, Vandenberghe J, et al. (2005) The impact of major depressive disorder on the short-and long-term outcome of Crohn's disease treatment with infliximab. Aliment Pharmacol Ther 22: 101–110. doi: 10.1111/j.1365-2036.2005.02535.x PMID: 16011668

50. Sanna L, Stuart AL, Berk M, Pasco JA, Girardi P, et al. (2013) Gastro oesophageal reflux disease (GORD)-related symptoms and its association with mood and anxiety disorders and psychological symptomology: a population-based study in women. BMC Psychiatry 13: 1.

51. Shim L, Talley NJ, Boyce P, Tennant C, Jones M, et al. (2013) Stool characteristics and colonic transit in irritable bowel syndrome: evaluation at two time points. Scand J Gastroenterol 48: 295–301. doi: 10.3109/00365521.2012.758767 PMID: 23320464

52. Crofts D, Michel VM, Rigby A, Tanner M, Hall D, et al. (1999) Assessment of stool colour in community management of prolonged jaundice in infancy. Acta Paediatr 88: 969–974. PMID: 10519339

53. Cryan JF, Dinan TG (2012) Mind-altering microorganisms: the impact of the gut microbiota on brain and behaviour. Nature reviews neuroscience 13: 701–712. doi: 10.1038/nrn3346 PMID: 22968153

54. Penninx BW, Milaneschi Y, Lamers F, Vogelzangs N (2013) Understanding the somatic consequences of depression: biological mechanisms and the role of depression symptom profile. BMC Med 11: 129. doi: 10.1186/1741-7015-11-129 PMID: 23672628

55. Maes M, Kubera M, Leunis JC, Berk M (2012) Increased IgA and IgM responses against gut commensals in chronic depression: Further evidence for increased bacterial translocation or leaky gut. Journal Of Affective Disorders 141: 55–62. doi: 10.1016/j.jad.2012.02.023 PMID: 22410503

56. Jacka FN, Cherbuin N, Anstey KJ, Butterworth P (2014) Dietary patterns and depressive symptoms over time: examining the relationships with socioeconomic position, health behaviours and cardiovascular risk. PLoS One 9: e87657. doi: 10.1371/journal.pone.0087657 PMID: 24489946

57. Dash S, Clarke G, Berk M, Jacka FN (2015) The gut microbiome and diet in psychiatry: focus on depression. Current opinion in psychiatry 28: 1–6. doi: 10.1097/YCO.0000000000000117 PMID: 25415497

58. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen Y-Y, et al. (2011) Linking long-term dietary patterns with gut microbial enterotypes. Science 334: 105–108. doi: 10.1126/science.1208344 PMID: 21885731

59. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, et al. (2014) Diet rapidly and reproducibly alters the human gut microbiome. Nature 505: 559–563. doi: 10.1038/nature12820 PMID: 24336217

60. Albenberg LG, Wu GD (2014) Diet and the intestinal microbiome: associations, functions, and implications for health and disease. Gastroenterology 146: 1564–1572. doi: 10.1053/j.gastro.2014.01.058 PMID: 24503132

61. Kim KA, Gu W, Lee IA, Joh EH, Kim DH (2012) High fat diet-induced gut microbiota exacerbates inflammation and obesity in mice via the TLR4 signaling pathway. PLoS One 7: e47713. doi: 10.1371/journal.pone.0047713 PMID: 23091640

62. Dipnall JF, Pasco JA, Meyer D, Berk M, Williams LJ, et al. (2015) The association between dietary patterns, diabetes and depression. J Affect Disord 174: 215–224. doi: 10.1016/j.jad.2014.11.030 PMID: 25527991

63. Olver JS, Hopwood MJ (2012) Depression and physical illness. Med J Aust 1: 9–12.

64. Shim RS, Baltrus P, Ye J, Rust G (2011) Prevalence, treatment, and control of depressive symptoms in the United States: results from the National Health and Nutrition Examination Survey (NHANES), 2005–2008. The Journal of the American Board of Family Medicine 24: 33–38. doi: 10.3122/jabfm.2011.01.100121 PMID: 21209342

65. Abbas OA (2008) Comparisons Between Data Clustering Algorithms. Int Arab J Inf Technol 5: 320–325.

66. Hagenaars JA, McCutcheon AL (2002) Applied latent class analysis: Cambridge University Press.

67. Eshghi A, Haughton D, Legrand P, Skaletsky M, Woolford S (2011) Identifying groups: A comparison of methodologies. Journal of Data Science 9: 271–292.

68. Freund Y, Schapire R, Abe N (1999) A short introduction to boosting. Journal-Japanese Society For Artificial Intelligence 14: 1612.

69. Natekin A, Knoll A (2013) Gradient boosting machines, a tutorial. Front Neurorobot 7.

70. Dipnall J, Pasco J, Berk M, Williams L, Dodd S, et al. (2017) Why so GLUMM? Detecting depression clusters through graphing lifestyle-environs using machine-learning methods (GLUMM). Eur Psychiatry 39: 40–50.